博士論文

Evaluation of Tree-based Models in Characterizing Oil and Gas Properties for Exploration and Production Processes
(探鉱・生産過程における石油および天然ガス特性解析向け木構造モデルの評価に関する研究)

指導教員森川博之教授

東京大学大学院工学系研究科

先端学際工学専攻

37-167332

アルマシャン メシャル ムサエド

Almashan Meshal Musaed

Abstract

This study is on the application of machine learning predictive models relating to upstream oil and gas. In particular, we build tree-based machine learning models to estimate the bubble point pressure (P_b) and the oil formation volume factor at the bubble point pressure (B_{ob}) of crude oils. Using the tree representation of the interpretable tree-based models, the feature importance and the predictions (decisions) made by the models can further be investigated heuristically, enhancing the understanding of models of reservoir characterization problems. In addition, the tree-based models are extended to predicting gas–liquid two-phase flow patterns and liquid holdup, which are considered challenging issues in producing gas wells. Moreover, in building our models, some of the datasets are novel. The novel datasets were collected from oil and gas wells in Kuwait. In addition, the trained models are evaluated in modeling incomplete and/or imbalanced datasets as real-world scenarios. Furthermore, comparative studies are made against other machine learning models and the upstream empirical correlations.

Machine learning has been successfully implemented for the past 20 years in estimating reservoir fluid properties competing with the empirical correlations. One of the most commonly utilized modeling schemes is the artificial neural network, which is known for its black-box problem in which the steps taken to reach the final estimation of the fluid properties are not shown. This study offers a different modeling approach that overcomes the limitations of the currently implemented modeling scheme, providing users with better predictions and a deeper understanding of the key input parameters in modeling. The evaluated models are tree-based models: the boosted decision tree regression (BDTR) module for regression, multiclass decision jungle (directed acyclic graphs; DAGs), decision tree, and random forest for classification.

Estimation of P_b and B_{ob}: Machine learning has been successfully implemented in the estimation of reservoir fluid properties, competing with the empirical correlations used in this field. One of the most commonly used modeling schemes is the artificial neural network, which is known for its black-box problem. This study offers a different modeling approach that overcomes this limitation. The model provides accurate estimations and facilitates a deeper understanding of the key input parameters and their importance to the estimated results. It uses a BDTR predictive modeling scheme to estimate the pressure–volume–temperature (PVT) properties, the bubble point pressure (P_b), and oil formation volume factor at the bubble point pressure (B_{ob}) as a function of oil and gas specific gravities, solution gas–oil ratio (R_s), and reservoir temperature. The robustness of the model is further evaluated by using incomplete datasets with imputation preprocessing. The built BDTR model exhibits higher accuracy and performance than previous machine learning models and the most commonly used empirical correlations for estimating P_b and B_{ob}.

***Estimation of liquid holdup in gas–liquid two-phase flow*:** In producing oil and gas wells, two-phase flow of gas and liquid is inevitable. Gas is less dense than liquid and travels faster as a result, leaving the liquid phase to build up in pipe segments. The amount of liquid occupying each pipe segment varies. This phenomenon is called liquid holdup, which is defined as the ratio of the liquid volume in a pipe element to the volume of the pipe element. Liquid holdup presents challenges in calculating mixture physical properties and two-phase flow pressure drop. Estimating liquid holdup is the first step in investigating production well problems. Early detection and identification of liquid holdup (H_L) in oil and gas wells are needed to reduce the maintenance downtime and thus increase production. This is crucial in designing

oil and gas facilities. Consequently, many studies on predicting liquid holdup have been carried out in the past and considering different flow conditions, resulting in a large number of empirical correlations with various degrees of accuracy.

Machine learning approaches to predicting liquid holdup in multiphase flows have been recently studied to improve the prediction accuracy compared to existing empirical correlations. However, these approaches ignore the heuristic feature importance of the input parameters to the predicted H_L values. In our study, a machine learning predictive model, BDTR, is trained, tested, and evaluated in predicating H_L in multiphase flows in oil and gas wells. Decision trees are considered as non-parametric machine learning models. The datasets used in training and testing the predictive model are experimental and were collected from literature (111 data points). Air–kerosene and air–water mixtures were used in obtaining the 111 experimental data points. The robustness of the model is further evaluated by using incomplete datasets with imputation preprocessing.

***Gas–liquid two-phase flow pattern classification*:** When gas and liquid are flowing simultaneously inside a transport conduit, their spatial distribution is referred to as the flow pattern of a multiphase flow. In the industry of petroleum engineering, multiphase flow characterization is required in many applications. Therefore, the industry is aiming for an accurate gas–liquid flow pattern predictive model in this field. Flow patterns change based on the gas and the liquid flow rates, the properties of fluid and gas, the diameter and the incline of the pipeline, and other fluid mechanical properties. The datasets used in this study were collected from a natural gas facility in Niigata, Japan, by the Japan Oil, Gas and Metals National Corporation (JOGMEC). The experimentally derived datasets include the superficial gas

velocity, the superficial liquid velocity, the flow pattern, the pressure, the temperature, and the liquid holdup. The data acquisition was performed with a pipeline of large diameter (106.3 mm) that was inclined at three different angles (0°, 1°, and 3°). The pressure was set at two different rates (592 and 2,060 kPa), and the liquid and gas flow rates covered a wide range of flow rates. In this study, a number of machine-learning-based models were trained and tested in predicting gas–liquid two-phase flow patterns. The multiclass decision jungle algorithm was used to build one of the predictive models. It consists of an ensemble of decision-directed acyclic graphs. The other machine learning algorithms evaluated in this study are the *k*-nearest neighbors, decision tree, and random forest multiclass classifiers. Since the dataset used in training the predictive models is imbalanced, over- and under-sampling techniques were used as a preprocessing step before training the predictive models. This step was applied to improve the prediction *accuracy* and *precision.* A comparative study is presented in this research between the studied models. Furthermore, the feature importance of the trained models is discussed.

***Evaluations of the proposed models*:** The evaluated models showed a superior performance, as the estimated values were in a good agreement with the experimental ones. In addition, the preprocessing steps for the incomplete datasets and the imbalanced datasets helped in improving the performance of the evaluated models. Moreover, the feature importance and the tree-representations of the evaluated models were discussed.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter provides an overview of the dissertation. I first sequentially discuss the research background, subject, and goal. Next, the main contributions of this dissertation are highlighted. Finally, the structure of this dissertation is elaborated at the end of this chapter.

1.1 Research background

The oil and gas industry is usually divided into three main sectors: upstream, midstream, and downstream. The upstream sector (or exploration and production; E&P) consists of operations and processes that are concerned with searching and exploring for underwater or underground natural gas and crude oil fields alongside drilling and operating the wells for producing natural gas and crude oil. The midstream sector includes crude or refined petroleum products transportation, storage, and wholesale marketing. The downstream sector includes purifying and processing natural gas and refining petroleum crude oil. The products that are derived from natural gas and crude oil are marketed and distributed in this sector.

In recent years, artificial intelligence has been considered for making significant changes in the energy sector, the oil and gas industry. In this research, we focus on the application of machine learning in the upstream sector as the most capital-intensive part of the oil and gas industry and a sector involving many uncertainties.

Many equations of state (EOSs) and empirical correlations have been developed for estimating the pressure–volume–temperature (PVT) properties of crude oils and are considered

as important parameters in many exploration and production processes. In addition, gas–liquid two-phase flow liquid holdup empirical correlations and flow-pattern mechanistic models have been developed to enhance the operation and production processes that involve gas–liquid multiphase flows. However, the developed PVT empirical correlations are accurate only if applied to oil samples collected from regions for which they were originally developed. Moreover, the gas–liquid two-phase flow liquid holdup empirical correlations and the flow pattern mechanistic models are not universal, and they are accurate only within the limited ranges that they were originally developed for; for example, horizontal or vertical orientations of the pipeline. Furthermore, EOSs are considered as computationally complex, and they require extensive details of the oil samples and the reservoir.

As a result, machine learning models have been recently studied in the oil and gas industry for more accurate and universal applications in upstream oil and gas processes. Fuzzy logic models, artificial neural networks, support vector machine regressions, and functional networks are among the studied models. However, compared to the mechanistic models, these machine learning models fail to express causality, which is considered a limited expressive capability. The upstream oil and gas sector is one of the sectors in which understanding is needed in order to improve the exploration and production processes through anomaly detection or preventive maintenance to reduce downtime and thus increase productivity.

In our study, we build tree-based machine learning models to estimate the PVT properties of crude oils. The studied PVT properties are the bubble point pressure (P_b) and oil formation volume factor at bubble point pressure (B_{ob}). Tree-based models are interpretable owing to their tree-representation. By the tree representation of the tree-based models, the feature importance and the predictions (decisions) made by the models can further be investigated heuristically, by

which the understandability of modeling the reservoir characterization problems can be enhanced. In addition, the tree-based models are extended to predict the gas–liquid two-phase flow patterns and the liquid holdup, which are considered challenging issues in producing gas wells. Some of the datasets used in building our models are novel. The novel datasets were collected from oil and gas wells in Kuwait. In addition, the trained models are evaluated in modeling incomplete and/or imbalanced datasets as real-world scenarios. Furthermore, comparative studies against other machine learning models, and the upstream empirical correlations, were carried out in our study.

In the following subsections, we further describe the background and problems that are related to PVT characterization and the gas–liquid two-phase flow-patterns and liquid holdup estimations.

1.1.1 Upstream oil and gas

As shown in **Figure 1**, the oil and gas industry is divided into three main sectors:

- Upstream: exploration and production.
- Midstream: transportation and processing.
- Downstream: refining and marketing.

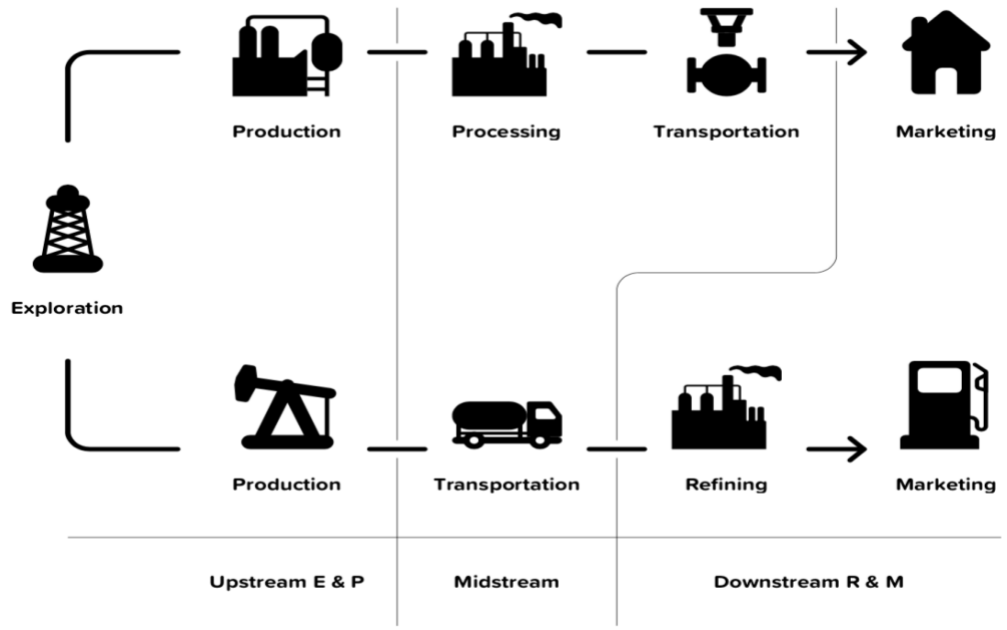

Figure 1—Oil and gas industry main sectors.

The focus of this research is upstream oil and gas (exploration and production). These are processes and operations concerned with searching and exploring for underground or underwater crude oil fields and natural gas and drilling and operating the wells to produce crude oil and natural gas.

Upstream oil and gas life cycle:

Licensing: Activities concerned with offering permission to manage a block area expected to contain natural gas and crude oil.

Exploration: Activities concerned with searching for natural gas and crude oil deposits on the reservoir underground and within the boundary of the block area.

Appraisal: Activities concerned with precisely defining the gas and crude oil characteristics and volume.

Development: Activities concerned with building the surface and subsurface facilities to produce natural gas and crude oil efficiently and safely.

Production: Activities concerned with extracting, processing, and exporting natural gas and crude oil as per contract agreement.

Abandonment: Activities concerned with plugging wells permanently, removing surface facilities, and restoring the block area as per initial state.

1.1.2 Problem—insufficient feature importance investigations

Non-tree-based models have limited understandability and expressive capabilities and an inability to determine the feature importance of the parameters used in conventional approaches such as empirical correlations. Understandability is a key issue in investigating producing oil and gas wells because of its benefits in reducing downtime and increasing productivity. Understandability can be enhanced by using interpretable models such as tree-based models, as in this study. The applications of machine learning models in industry domains are limited if they lack interpretability (Rudin, 2019). Interpretability will help in better understanding PVT characterization and will help in investigating the gas–liquid two-phase flow liquid holdup and flow pattern problems. This can be achieved by referring to the heuristic feature importance generated by the tree-based models and by referring to the determined permutation feature importance.

1.2 Research goal

This study aims for a comprehensive and intensive evaluation of tree-based models for the estimation of P_b and B_{ob}, gas–liquid two-phase flow pattern classification and estimation of liquid holdup in gas–liquid two-phase flow. The intensive evaluation includes preprocessing data imputation and over- and under-sampling.

The key questions we aim to answer are the following:

Estimation of P_b and B_{ob}:

- Can we build an accurate BDTR model for multiregional and regional datasets?
- Which parameter is most essential in estimating the P_b and B_{ob} by a BDTR model?

Estimation of liquid holdup in gas–liquid two-phase flow:

- Is the BDTR model that is accurate for P_b and B_{ob} also accurate for estimating liquid holdup in gas–liquid two-phase flow?
- Which parameter is most essential in estimating liquid holdup by a BDTR model?

Gas–liquid two-phase flow pattern classification:

- Is the multiclass decision jungle (directed acyclic graphs; DAGs) model accurate in classifying gas–liquid two-phase flow patterns?
- Can over- and under-sampling preprocessing improve the accuracy of the decision tree, random forest, and DAGs models in classifying gas–liquid two-phase flow patterns?
- Which parameter is most essential in classifying the flow pattern for high and low pressure by the DAGs model?

1.3 Main contributions

The main contributions of the presented research are summarized as follows:

Estimation of P_b and B_{ob}:

- Superior performance on both multiregional and regional datasets (complete datasets)
- Robust even with the existence of missing data points in the datasets (incomplete datasets) and on unseen datasets
- Determination of the most essential parameter (feature)

Estimation of liquid holdup in gas–liquid two-phase flow:

- Superior performance in estimating H_L
- Robust even with the existence of missing data points in the dataset (incomplete dataset)
- Determination of the most essential parameter (feature)

Gas–liquid two-phase flow pattern classification:

- Superior performance of the decision tree and DAGs models in classifying gas–liquid two-phase flow patterns
- Over- and under-sampling preprocessing showed improvement in the performance of the evaluated models: decision tree, random forest, and DAGs.
- Determination of the most essential parameter (feature) for the DAGs model

1.4 Thesis structure

In this section, we review the structure of this dissertation.

In Chapter 2, we first give a detailed introduction to related works of PVT characterization. Specifically, we discuss the empirical correlations developed in the past for estimating the PVT properties. After, we discuss the boosted decision tree regression model. Later, we describe the data acquisition of the datasets that were used in training and testing our predictive models. In addition, in this chapter we present the experimental setup and the hyperparameter settings of our study for estimating P_b and B_{ob}. Moreover, in this chapter we evaluate the developed models by first explaining the evaluation metrics and then presenting a comparative study against empirical correlations and other machine learning models. Furthermore, the permutation feature importance (PFI) is discussed in this chapter. Lastly, this chapter concludes the evaluation study of the estimation of P_b and B_{ob}.

In Chapter 3, we first give a detailed introduction to related works of estimating liquid holdup in gas-liquid two-phase flow. Specifically, we discuss the empirical correlations developed in the past for estimating the liquid holdup. After, we describe the data acquisition of the datasets that were used in training and testing our predictive models. In addition, in this chapter we present the experimental setup and the hyperparameter settings of our study for estimating liquid holdup. Moreover, in this chapter we evaluate the developed models by first explaining the evaluation metrics and then presenting a comparative study against empirical correlations and other machine learning models. Furthermore, the permutation feature importance (PFI) is discussed in this chapter. Lastly, this chapter concludes the evaluation study of the estimation liquid holdup in gas-liquid two-phase flow.

In Chapter 4, we first give a detailed introduction to related works gas–liquid two-phase flow flow-patterns classification. Specifically, we discuss the related work of the gas–liquid two-phase flow pattern classification that was conducted by Japan Oil, Gas and Metals National Corporation (JOGMEC). After, we describe the data acquisition of the datasets that were used in training and testing our predictive models. In addition, in this chapter we present the experimental setup and the hyperparameter settings of our study for classifying flow patterns. Moreover, in this chapter we evaluate the developed models by first explaining the evaluation metrics and then presenting a comparative study against empirical other machine learning models. Furthermore, the permutation feature importance (PFI) is discussed in this chapter. Lastly, this chapter concludes the evaluation study of classifying flow-patterns for gas-liquid two-phase flow.

In Chapter 5, we conclude this dissertation and point out potential future research directions.

Chapter 2

Estimation of P_b and B_{ob}

2.1 Introduction

Accurate and reliable PVT data are crucial in the chemical and petroleum engineering field, specifically for material balance calculations, numerical reservoir simulations, reserve estimates, and inflow performance characterization (De Ghetto *et al.*, 1995). For upstream oil and gas processes, the more accurate the PVT properties, the better the results and performance of their applications, such as in oil and gas reservoir engineering, development, and production. This, in turn, saves costs, time, and effort for process designs and optimizations.

The accurate and precise estimations of P_b and B_{ob} are important, as they are the most influential PVT properties affecting oil reserves and well production. There are three commonly used methods to determine P_b and B_{ob}: EOSs, empirical PVT correlations, and artificial neural networks (ANNs). EOSs are based on detailed experimental data on the reservoir fluids' composition (Fath *et al.*, 2018). These experimental data are not easily obtained from field parameters, as the process of acquiring them is costly and time consuming. Therefore, such experimental data are not always available (Fath *et al.*, 2018). Empirical PVT correlations, however, are based on easily observed experimental data, such as reservoir temperature and pressure, as well as the specific gravity of gas and oil.

The accuracy of empirical correlations has been investigated by numerous authors, who have shown that the accuracy of these correlations chiefly depends on the characteristics of the

datasets (De Ghetto *et al.*, 1995). Highly accurate estimates are obtained in correlations performed with "regional" datasets, which are acquired from the same geographic location (De Ghetto *et al.*, 1995; Fath *et al.*, 2018). By contrast, "multi-regional" datasets acquired from different locations do not work well for reliable estimates of PVT properties.

ANNs are accurate machine learning models that have been tested in the literature on regional and multiregional datasets (Ghrabi *et al.*, 2003). However, ANNs are black-box models because their internal decision-making mechanisms are without interpretable representation (El-Sebakhy *et al.*, 2009). The applications of machine learning models in industry domains are therefore limited (Rudin, 2019). With the help of interpretability, the developers of any interpretable machine learning model can interrogate why a machine learning predictive model behaves in a certain way and, consequently, develop improvements. Interpretability is also needed in the upstream oil and gas industry. Interpretability in upstream oil and gas can help specify the key parameters, which can further help in investigating the physical meaning behind the estimations made by the models. Physical meaning is the explanation of results in a physical and scientific way. Deriving physical meaning helps in understanding why and how these results are generated for certain sets of features and settings.

From these viewpoints, we focused on BDTRs to estimate the PVT properties of crude oil systems. BDTRs are decision tree-based models, which are interpretable owing to tree representation. The main characteristic of the BDTR model is that it is based on boosting, which refers to a series of decision trees built in a stepwise fashion, followed by aggregating the results from each step. Boosting improves the performance of decision tree ensembles. The key questions we address here are the following: "Can we build an accurate BDTR model for multi-

regional and regional datasets?" and "Which parameter is the most essential one in estimating the PVT properties by a BDTR model?"

To answer these questions, we built and intensively investigated the performance of the BDTR model. Our analysis showed that the BDTR model outperformed the most commonly used empirical correlations—ANNs and random forests—in terms of accuracy in estimating P_b and B_{ob}. The analysis of the BDTR model in terms of feature importance revealed the solution gas–oil ratio (R_s) to be the most essential input parameter for the model. The contributions of this study are as follows.

First, the BDTR model showed superior performance on a multiregional dataset. The dataset was collected from oil wells in different geographic regions representing 350 different crudes from major oil fields in the North Sea, Middle East, North America, South America, Africa, South-East Asia, and other parts of the world. This implies that the BDTR model is universal and can therefore be applied to datasets with diverse crude oil data from all over the world.

Second, the BDTR model showed superior performance not only on multiregional datasets but also on regional datasets. The regional dataset was collected from the Greater Burgan oil field, the largest oil field in Kuwait and the world's largest sandstone oil field.

Third, we showed that the BDTR model's performance is robust, even with the existence of missing data points in the dataset, by testing the accuracy of the model on datasets with multiple missing data points as an incomplete dataset.

Fourth, we showed the robustness of the BDTR model against an unseen dataset, which is the test data that was not included in the training dataset. This yielded an effective and

accurate multiregional dataset-trained BDTR model that could be used to estimate unseen and regional datasets. This also showed that the model does not suffer from overfitting.

Fifth, we analyzed the BDTR model in terms of feature importance based on two approaches: the PFI algorithm and tree-representation analysis. Even though these approaches are independent of each other, our analysis revealed that either approach determined R_s to be the most essential input parameter for the BDTR model in estimating P_b and B_{ob}.

2.2 Literature review

In this section, previously developed laboratory analysis, EOSs, empirical correlations, and machine learning models for estimating PVT properties, such as P_b and B_{ob}, are reviewed as related work to further highlight the importance of the current study.

In hydrocarbons, the pressure at which the first bubble of gas comes out of an oil solution is referred to as the P_b. At the P_b, the ratio of the volume of oil in reservoir conditions to the volume in stock-tank conditions is referred to as the B_{ob}. These two parameters are among the most important PVT properties of oil and gas systems. They are crucial in reservoir and production engineering calculations, such as reservoir performance estimations, well and reservoir simulations, mass balance calculations, production facility designs, flow performance calculations, economic evaluations, and oil recovery enhancement projects (El-Sebakhy, 2009; McCain *et al*., 1998; Rafiee-Taghanaki *et al.*, 2013).

The P_b and B_{ob} can be determined through constant composition expansion laboratory tests of reservoir fluid samples (Ahmed, 1989). In such tests, at a pressure initially higher than that of the reservoir pressure and at a temperature similar to that of the reservoir, a reservoir fluid sample is first placed in a visual PVT cell. Next, the pressure is gradually reduced, while the temperature is maintained at a constant level. A plot of the volume of the hydrocarbon sample

against the pressure is drawn. P_b is then determined as the point at which the slope starts to change (Ahmed, 1989). Laboratory analysis is a very accurate method for determining P_b and B_{ob}. However, it requires time, money, and effort, in addition to extra care in handling oil reservoir fluid samples (McCain, 1990).

As an alternative approach, EOSs can be used to estimate PVT properties. However, EOSs are complex and require large amounts of reservoir fluid composition data. Consequently, several empirical correlations have been developed to estimate PVT properties. Empirical correlation calculations are much simpler than EOS calculations. For accuracy, some of these correlations have been developed for crude oil systems using regional datasets of certain geographic locations and with specific types of chemical composition. The main techniques for developing these correlations are graphical, linear, and nonlinear regression. The major assumption is that P_b and B_{ob} are functions of reservoir temperature (T), R_s, oil-specific gravity (γ_o), and gas specific gravity (γ_g). Examples are works by De Ghetto *et al.* (1995), Kartoatmodjo *et al.* (1994), Standing (1947), Vazquez and Beggs (1980), Glasø (1980), Al-Marhoun (1980), Labedi (1990), Dokla *et al.* (1992), Petrosky and Farshad (1993; 1998), Dindoruk *et al.* (2004), Lasater (1958), Omar *et al.* (1993), Naseri *et al.* (2005), Khairy *et al.* (1998), Macary *et al.* (1993), Frashad *et al.* (1996), Almehaideb (1997), Al-Shammasi (2001), and Jarrahian *et al.* (2015).

However, such empirical correlations have varying degrees of accuracy, as they allow for only limited ranges of PVT input features as regional datasets. Further, most of these correlations have not produced reliable results (Fath, Pouranfard and Foroughizadeh, 2018). Consequently, several machine-learning approaches have been used for PVT characterization to increase the accuracy of the estimations of crude oil PVT properties.

Many researchers have implemented ANNs as an alternative reliable solution for estimating the PVT properties of crude oils. Feedforward neural networks and back-propagation (BP) training algorithms are the most commonly used methods for modeling ANNs. Gharbi and Elsharkawy (1997) developed two models: one for estimating the B_{ob} and the other for estimating the P_b based on neural networks. The models were built with regional datasets of crude oil systems in the Middle East. Gharbi and Elsharkawy's models generated lower standard deviations and relative errors than the most commonly used empirical correlations for estimating the B_{ob} and P_b. Further, Gharbi and Elsharkawy (2003) proposed another ANN model for estimating P_b and B_{ob} of crude oils from worldwide (multiregional) datasets, which is the subject of the main comparative study in this paper. However, some studies have not adopted ANN techniques because they represent black-box modeling schemes. They also suffer from the extrapolation problem.

Using a radial basis function neural network model, Elsharkawy (1998) proposed a new method for estimating oil viscosity, B_o, saturated oil density, undersaturated oil compressibility, R_s, and evolved gas. Osman *et al.* (2001) developed a neural network model with a 4-5-1 multilayer feedforward structure and a BP algorithm to estimate the B_{ob}.

Malallah *et al.* (2006) estimated B_o and P_b by using the alternating conditional expectation algorithm. A multiregional dataset with 5,200 data points for oil samples collected from oil fields around the world (e.g., North and South America, the North Sea, Southeast Asia, the Middle East, and Africa) was used in the development of this model. Moghadassi *et al.* (2009) developed a model to estimate PVT properties based on ANNs. Hashemi Fath (2018) developed a multilayer feedforward neural network model for estimating the P_b of crude oils. Based on previous studies, a multiregional dataset of 760 data points for crude oil samples

collected from oil fields around the world was used to train and test the model. Ramirez *et al.* (2017, p. 34) proposed a mathematical model using machine learning for estimating bubble point pressure and oil formation volume factor. The proposed model consists of principal component analysis (PCA) and an ANN. PCA is used for data decorrelation to reduce redundancy in the dataset, which in return reduces the number of neurons and hidden layers in the ANN. Their dataset consists of 504 data points from Malaysia, the North Sea, and the Middle East fields. Olatunji *et al.* (2015, p. 35) demonstrated the power of a type-2 fuzzy logic system (FLS) in improving the estimation of PVT properties and permeability in a hybrid intelligent model setup that consists of a type-2 FLS (T2) and a sensitivity-based linear learning method (SBLLM). Type-2 FLS is used for dealing with the uncertainties existing in the datasets and is aimed at improving the performance of the subsequent method in the hybrid model, the SBLLM.

As only a few studies have considered evaluating their models on multiregional datasets, the present study aimed to build an accurate universal (for worldwide crude oils) predictive model for estimating the P_b and B_{ob} of crude oils based on a BDTR algorithm. Therefore, a dataset of 5,207 data points was used to build and evaluate the model. The multiregional dataset was derived from oil samples from a wide range of crude oils from different geographic locations around the world (Gharbi *et al.*, 2003). The performance of the BDTR model was compared with ANNs and the most commonly used empirical correlations. The importance of each input parameter for the BDTR model used in estimating P_b and B_{ob} was determined.

In our related work (Almashan *et al.*, 2019a), *k*-means clustering was evaluated in estimating P_b and B_{ob} for worldwide datasets as a preprocessing technique for the BDTR model. The comparative study revealed that *k*-means clustering is insignificant in our approach. Therefore, it was not considered in this extended version of our work. In addition, the present

work evaluates the BDTR model for estimating P_b and B_{ob} for incomplete multiregional datasets, which was not considered in our previous work; in that work (Almashan *et al*., 2019b), incomplete datasets were considered for the Kuwait regional dataset. The complete Kuwaiti regional dataset is considered in our extended work presented in this paper. In addition, an evaluation of the robustness of the model trained on the worldwide multiregional dataset being tested on the Kuwaiti regional dataset, as an unseen dataset, is first presented in the current work. Moreover, the feature importance is extended and further discussed in our present work for both datasets: the multiregional and the regional datasets. Therefore, this work is an extension and an improvement of our previous related works.

Table 1 shows the related works for estimating P_b and B_{ob}.

Table 1—Related works for estimating P_b and B_{ob}.

Approach	Pros	Cons	Applicability	Interpretability
EOSs	More accurate compared to the empirical correlations	Computationally complex	Universal if tuned Regional	No
Empirical correlations	Less complex	Lower accuracy compared to EOS Limited to certain regions	Not universal (not multiregional) Regional	No
Artificial neural networks, support vector machine, Functional Network, other non-tree-based models	Accurate Not complex	Less accurate compared to the laboratory analysis Lack of interpretability	Universal if trained with multiregional datasets Regional	No
Boosted decision tree regression (This research)	More accurate Not complex Heuristic feature importance Interpretability	Less accurate compared to the laboratory analysis	Universal if trained with multiregional datasets Regional	Yes

2.3 Boosted decision tree regression and feature importance

2.3.1 Modeling of boosted decision tree regression

In this subsection, the BDTR model is explained. The details of the modeling and hyperparameter settings of the BDTR model are provided.

Boosting is an ensemble learning technique for minimizing training errors by combining a group of weak learners into a strong learner. A random sample of data is chosen, fitted with a model, and then trained in a sequential manner in boosting—that is, each model attempts to compensate for the shortcomings of its predecessor.

In the BDTR model, each constructed decision tree is dependent on prior trees. The algorithm learns to fit the residual of the preceding trees, which constitutes boosting. The boosting algorithm utilized in the training of the proposed model was the multiple additive regression tree (MART) gradient-boosting algorithm. This algorithm sequentially combines weak learners in such a way that each new learner fits the previous step's residuals, thereby improving the model. The final model aggregates the results from each step, resulting in a strong learner, and this learner was then used to estimate future P_b or B_{ob} values. Decision trees are used as weak learners in the gradient boosted decision tree algorithm in the BDTR model. The mean absolute error (MAE) is used in our model as the loss function for determining the residuals.

In the BDTR model, for each node in the decision tree, the data are split into two groups by finding the threshold that generates the smallest MAE. For choosing a new node or a root, the MAE for each node candidate is compared, and the one with the minimum MAE is chosen as the new node.

Tree-based models including BDTR can find non-linear relationships as in our dataset, that is why it was chosen over linear regression. For P_b and B_{ob} non-linearity exists between the dependent variable and all the independent variables except for R_s.

The Azure machine learning platform was used for training and testing the predictive models with the Boosted Decision Tree Regression module (Microsoft, 2019a), which is based on FastTree library on ML.NET. FastTree is an efficient implementation of the MART gradient boosting algorithm.

A BDTR model was trained and tested for the worldwide dataset. Two models were built: one for estimating P_b and the other for estimating B_{ob}. The input features of the models are R_s, T, γ_{API}, and γ_g. Despite the slight possibility of less coverage, boosting is used to improve the accuracy of decision tree ensembles. In this study, the boosted decision tree model was used for regression as a supervised learning method in which the label column is a numerical value (i.e., P_b or B_{ob}). We also trained, validated, and tested a third predictive BDTR model for estimating P_b values for Kuwaiti crude oil systems.

$$P_b = f_1(T, \gamma_{API}, R_s, \gamma_g). \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$B_{ob} = f_2(T, \gamma_{API}, R_s, \gamma_g). \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Datasets gathered from different geographic locations require specific hyperparameter settings to improve the *accuracy* and *precision* of the model. A dataset of worldwide crude oils with wide ranges for each of the input features (**Table A-1**) requires accurate and precise tuning of the model hyperparameters. Therefore, tuning the hyperparameters of the BDTR algorithm to generate the best estimation results was one of the main tasks in modeling PVT properties in this

study. The process of tuning the hyperparameters—or hyperparameter optimization—of each trained model was performed by first performing a random search and then retraining the model using a grid search for a more specific set of hyperparameters based on the results generated for the random search.

The number-of-decision-trees hyperparameter indicates the total number of decision trees to build in the ensemble. Theoretically, better coverage can be achieved by making more decision trees, but training time increases. The number-of-decision-trees hyperparameter is equivalent to the number of iterations or the number of estimators in the model, as each decision tree takes a single iteration in the learning process. The maximum depth of any decision tree can be restricted by the maximum depth of the decision tree's hyperparameter. Increasing the tree's depth can improve precision but at the cost of some overfitting and more training time.

The minimum number of cases needed to establish any terminal node (leaf) in a tree is indicated by the min-number-of-samples-per-leaf-node hyperparameter. The threshold for developing new rules is raised by increasing this value. With a value of 1, even a single case will result in the creation of a new rule. If the value is increased to 4, the training data must include at least four cases that follow the same criteria.

The number of splits to use when constructing each node of the tree is indicated by the number-of-random-splits-per-node hyperparameter. A split denotes a random division of features at each level of the tree (node). The learning-rate hyperparameter specifies the step size; the value indicated must be between 0 and 1. This hyperparameter controls the pace at which the model converges on the optimal solution, either fast or slow. A larger value indicates a larger step and faster convergence at the risk of overshooting the optimal solution. A smaller value indicates a smaller step and a slower convergence at the risk of extending the training time.

Different numbers of learning rates were tested for each model, and the final results were compared for their convergence to the optimal solution. In this comparison of the trained models, the model with the lowest mean absolute error was chosen as the best-trained model.

The max-number-of-leaves-per-tree hyperparameter specifies the maximum number of terminal nodes (leaves) that can be created on a tree. At the risk of longer training times and overfitting, increasing the value of this hyperparameter will potentially increase the tree's size and consequently achieve better precision. To avoid overfitting, this parameter has been maintained to be equal to or less than a predefined threshold.

Figure 2 shows the boosted decision tree regression algorithm flowchart.

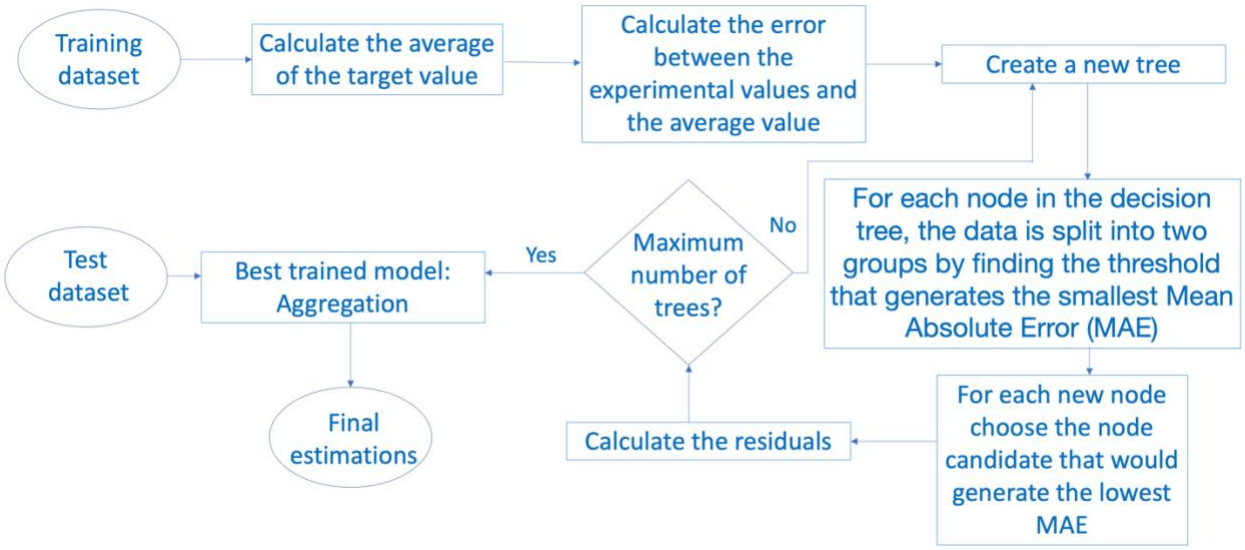

Figure 2—Boosted decision tree regression algorithm flowchart.

2.3.2 Feature importance determination

The feature importance of the input parameters for the trained predictive model was determined using the PFI algorithm. This algorithm determines the importance scores of the input feature variables for a given dataset. The importance score of the corresponding feature is

determined based on the model's sensitivity, which is measured by randomly permuting the values of the input features. By choosing an evaluation metric, the deviation of that metric after the permutation of the values of a certain feature quantifies the contribution or importance of that particular input feature of the model to the performance of the estimation of the target feature, which is P_b or B_{ob} in this case. If permuting the values of an input feature results in a considerable reduction in the predictive power of the model, this feature is important in estimating the output feature. The Azure machine learning platform was used for determining feature importance with the PFI module based on the permutation importance algorithm (Microsoft, 2019d).

By referring to the model's tree representations in the gradient-boosting ensemble, if most of the node splits, including the root node, were based on a certain feature, then this feature has a higher importance for the model. For these node splits, the method of splitting the nodes was based on choosing the threshold that would generate the lowest MAE, where each new node or root was chosen if it has the lowest MAE from the node candidates. This implies that splitting the nodes based on that feature can reduce MAE in the learning process and thus generate more accurate estimations of the target values.

Figure 3 shows the permutation feature importance algorithm flowchart.

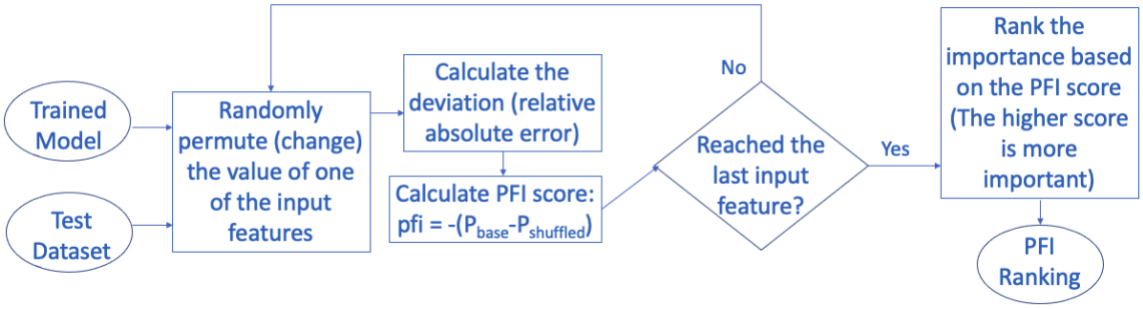

Figure 3—Permutation feature importance algorithm flowchart.

2.4 Data acquisition

In this section, the data acquisition-related information is detailed for both the worldwide (multiregional) dataset and the Kuwaiti crude oil (regional) dataset. We evaluated the model based on two different datasets: worldwide and regional datasets. The worldwide dataset covers a wide range of crude oils from all over the world, with wider ranges of parameters compared to the regional dataset. The regional dataset represents Kuwaiti crude oil systems and has not been previously studied. The universal model trained on the worldwide dataset was also tested on the regional Kuwaiti dataset as an additional validation step for the performance of the BDTR model.

2.4.1 Worldwide dataset (multiregional dataset)

The worldwide crude oil dataset consisted of 5,207 data points. The data points were collected from oil wells in different geographic regions representing 350 different crudes from major oil fields in the North Sea, Middle East, North America, South America, Africa, South-East Asia, and other parts of the world (Gharbi *et al.*, 2003). The dataset consists of the PVT properties: P_b, B_{ob}, T, R_s, γ_g, and γ_{API}. The statistical measures of the PVT datasets are shown in Table A-1. The datasets cover a wide range of values representing diverse crude oil samples, for which the P_b values range from 79 psi to 7,130 psi, the B_{ob} values range from 1.02 bbl/stb to 2.92 bbl/stb, the T values range from 74°F to 342°F, the R_s values range from 9 scf/stb to 3370 scf/stb, the γ_g values range from 0.5 to 1.67, and the γ_{API} values range from 14.3 to 59. Approximately 70% of the data points in the dataset were used for training, and 30% were used for validation and testing.

Cross plots that show the T, R_s, γ_g, and γ_{API} against the experimental bubble point pressure for all data are shown in **Figure A-1**, **Figure A-2**, **Figure A-3**, and **Figure A-4**, respectively. Cross plots that show the T, R_s, γ_g, and γ_{API} against the experimental oil formation volume factor at bubble point pressure for all data are shown in **Figure A-5**, **Figure A-6**, **Figure A-7**, and **Figure A-8**, respectively**.**

2.4.2 Kuwaiti crude oil dataset (regional dataset)

In this study, Kuwaiti crude oil systems were considered as a regional dataset to test and compare the predictive power of the BDTR model against the most commonly used empirical correlations that are most commonly used for Kuwaiti crude oil systems. This dataset is new and was used for the first time in the literature. It consists of 117 data points derived from the Greater Burgan oil fields representing Kuwaiti crude oil systems. The dataset consists of the PVT properties P_b, T, R_s, γ_g, and γ_{API}. B_{ob} is not present in the Kuwaiti regional dataset, as the acquired PVT report did not include it. The statistical measures of the PVT datasets are shown in **Table A-2**. The T values range from 112°F to 143°F, the R_s values range from 45 scf/stb to 701 scf/stb, the γ_g values range from 0.843 to 1.227, the γ_{API} values range from 9.68 to 41.5, and the P_b values range from 313.73 psi to 1,940 psi. The T, Rs, γ_{API}, and γ_g are the input features used in the BDTR predictive model. The P_b is the output feature. Cross plots that show the T, R_s, γ_g and γ_{API} against the experimental bubble point pressure for all data are shown in **Figure A-9, Figure A-10, Figure A-11,** and **Figure A-12,** respectively. The raw dataset is shown in **Table A-5**.

2.5. Experimental setup

- Experimental work was performed on the Azure machine learning platform.
- Boosted Decision Tree Regression module for the BDTR model.
- Permutation Feature Importance module for the PFI scores.
- ML.NET machine learning library for C# programming language (Machine Learning Studio).

The setting of the hyperparameters proceeded as follows:

1. Hyperparameter settings started with a random search followed by a grid search (a limited combination).
2. Fine-tuning was applied to the hyperparameters to obtain the best model.
3. Different numbers of learning rates were tested for each model, and the results were compared for their convergence to the best model.
4. To avoid overfitting, the max-number-of-leaves-per-tree hyperparameter was maintained below the threshold:

   $TH\_max\_num\_leaves \leq 60\%$ training dataset.

Incomplete multiregional (worldwide) dataset:

- Data imputation was applied to 18% of the total test dataset for each of the input features with missing values.
- The data imputation was based on the average value of each of the corresponding features in the test dataset.

Incomplete Kuwaiti crude oil system dataset (the regional dataset):

- Data imputation is applied to 25% of the total test dataset for each of the input features with missing values.
- The data imputation is based on the average value of each of the corresponding features in the entire dataset.

Incomplete regional and multiregional datasets:

- The imputed missing data points were included in the test dataset only. The training dataset is complete.
- Missing data points are data points that were removed intentionally and randomly from the dataset, covering a wide range of the data values.
- The number of the missing data points was chosen randomly, as missing data points occur randomly in real-life scenarios.
- For the regional dataset, 25% was chosen as the dataset is smaller compared to the multiregional dataset, which was set at 18%.

Unseen dataset:

- The robustness of the universal model (multiregional BDTR) is tested on an unseen, incomplete, and regional dataset to validate the model's applicability and to show that it does not suffer from overfitting.
- The dataset used for training is not the same dataset used for testing:
    - ❑ Training: complete multiregional dataset

- ❑ Testing: incomplete regional dataset (117 data points and 48 data points separately)

2.6. Evaluation methodology

2.6.1 Evaluation criteria

The following evaluation criteria were considered to comprehensively assess the BDTR model when applied to multiregional, regional, and incomplete datasets. The robustness of the universal model, the model that was trained on a worldwide (multiregional) dataset, was tested on an unseen, incomplete, and regional dataset to validate the model's applicability. Furthermore, the feature importance of the input parameters is considered in the evaluation study to better understand the model's estimation process.

(a) Evaluation of the complete multiregional dataset

The goal of this evaluation was to compare the performance of the BDTR model with empirical correlations when applied to complete multiregional datasets. We tested the overall performance of the BDTR model trained and tested on the multiregional dataset, as a complete dataset with no missing values. The model was compared to ANNs, random forest regressions, and the most commonly used empirical correlations for estimating P_b and B_{ob}. Furthermore, the PFI of the BDTR model was determined.

(b) Evaluation of the complete Kuwaiti crude oil dataset

The goal of this evaluation was to assess the BDTR model's performance against empirical correlations when applied to a complete Kuwaiti crude oil regional dataset. We assessed the overall performance of the BDTR model trained and tested on the Kuwaiti crude oil (regional) dataset as a complete dataset with no missing values. The model was compared to the

most commonly used empirical correlations for the Kuwaiti crude oil dataset. Furthermore, the PFI of the BDTR model was determined.

(c) Evaluation of incomplete multiregional and Kuwaiti crude oil datasets

This evaluation assessed the performance of the BDTR model's performance when applied to incomplete multiregional or regional datasets. The performance and predictive power of a machine learning model greatly rely on the quality and completeness of the dataset used in building the model. Data imputation is a technique in which missing data are substituted with values to improve the accuracy of the model and avoid bias that might be introduced by the missing values.

To evaluate the performance and robustness of the BDTR model, the model was retrained and tested on the multiregional dataset as an incomplete dataset in which multiple missing data points exist in the test data by intentionally removing them from the dataset. The average value of each input feature was used to replace the corresponding missing values. This was also applied to the regional dataset: the Kuwaiti crude oil dataset.

Regarding the incomplete multiregional (worldwide) dataset, data imputation was applied to 18% of the total test dataset for each of the input features that had missing values. The data imputation was based on the average value of each of the corresponding features in the test dataset. The training dataset is complete, with no missing data points.

For the Kuwaiti crude oil (regional) dataset, each input property in the dataset used in training and validating the BDTR model was 100% complete. However, 48 data points of the total dataset were used in testing the predictive model as an unseen dataset for the model, with each of the input features containing 12 missing data points and imputation applied based on the average values of the corresponding properties for the whole dataset excluding the missing points; 105 data points in total were used for calculating the average value.

(d) Evaluation of the universal model on incomplete Kuwaiti crude oil datasets

We assessed whether the multiregional BDTR model's performance would remain high even when the model was tested on unseen, incomplete, and imputed regional datasets. The estimation accuracy of the universal model, trained on the complete multiregional dataset, was tested on the unseen regional and incomplete Kuwaiti crude oil dataset. Although the other scenarios in this study involved dividing the same dataset for training, validation, and testing, in this scenario, the dataset used for training was not the same dataset used for testing.

(e) Feature importance

We determined the feature importance of each of the input parameters in estimating the target values, P_b and B_{ob}, for both the regional and multiregional datasets. The feature importance of the input parameters for the trained predictive model was determined using the PFI algorithm. This was studied for both the multiregional and regional datasets. The model's tree representation was examined to further validate the feature importance generated by the PFI algorithm.

2.6.2 Evaluation measures

The estimation accuracy of the built models was evaluated using the following accuracy measures, which are the most commonly used error metrics in evaluating the performance of regression models:

Mean absolute percentage error (MAPE: E_a)

The relative absolute deviation between the experimental and estimated values was determined using this statistical accuracy measure. A lower E_a% value indicates higher accuracy.

It is defined as follows:

$$E_a\% = \frac{1}{n}\sum_{i=1}^{n} |E_i\%| \,. \quad \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots (3)$$

where E_i% denotes the relative deviation between the experimental and estimated values.

$$E_i\% = \left(\frac{x_{exp} - x_{pred}}{x_{exp}}\right)_i \times 100 \quad i = 1, 2, \ldots, n. \quad \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots (4)$$

where x_{exp} and x_{pred} denote the experimental and estimated values, respectively. The total number of data points is denoted by n.

2.7. Results and discussion

2.7.1. Performance of the multiregional dataset

The multiregional dataset was considered in this study to evaluate the performance of the BDTR model as a universal model in estimating P_b and B_{ob} for crude oil samples from all over the world. The results indicate that the BDTR model outperformed the most commonly used empirical correlations in estimating P_b and B_{ob} in terms of E_a. A lower value of E_a indicates a higher correlation between the estimated and experimental values. The studied empirical correlations included Glasø (1980), Dokla and Osman (1992), Petrosky *et al.* (1993), Al-Marhoun (1988), Frashad *et al.* (1996), Vasquez and Beggs (1980), Kartoatmodjo *et al.* (1994), and Standing (1947). The performance was also compared to that of the ANN predictive model. The structure of the ANN model consists of one hidden layer with five neurons (Gharbi *et al.*, 2003). The statistical results of the comparative study are presented in **Table 2** (for P_b) and **Table 3** (for B_{ob}). The tables show that the BDTR predictive model has higher accuracy than

ANNs, the random forest regression model, and the most commonly used empirical correlations in estimating P_b and B_{ob}.

The E_a of the tested correlations and models for estimating P_b are as follows: 27.82 for Labedi's correlation, 28.17 for Kartoatmodjo's correlation, 32.98 for Vasquez and Begg's correlation, 34.32 for Glasø's correlation, 39.82 for Dokla and Osman's correlation, 25.20 for Standing's correlation, 49.61 for Al-Marhoun's correlation, 22.73 for Frashad's correlation.,15.38 for the ANN model (Gharbi *et al.,* 2003), 14.13 for the random forest regression, and 9.17 for the BDTR model (complete).

The E_a of the tested correlations and models for estimating B_{ob} are as follows: 2.38 for Frashad's correlation, 2.53 for Abdul Majeed's correlation, 2.10 for Kartoatmodjo's correlation, 3.51 for Vasquez and Beggs's correlation, 4.32 for Glasø's correlation, 3.21 for Dokla and Osman's correlation, 2.47 for Standing's correlation, 2.35 for Al-Marhoun's correlation, 2.52 for Petrosky's correlation, 2.62 for Labedi's correlation, 2.49 for Obomanu's correlation, 3.40 for Elsharkawy correlation, 2.04 for the ANN model (Gharbi *et al.,* 2003), 1.34 for random forest regression, and 0.85 for the BDTR model (complete).

As shown in Table 2 and Table 3, the estimation accuracy of B_{ob} was higher than that of P_b. This is due to B_{ob} having a limited set of values, as opposed to their absolute values. This is in contrast to those of P_b, which had a much broader spectrum in terms of absolute values.

A cross plot of the estimated P_b values to the corresponding experimental ones is shown in **Figure 4**. The corresponding estimated B_{ob} values for the experimental data are presented in **Figure 5**. As shown in these figures, the estimated values generated by the built BDTR model are in good agreement with the experimental values.

The hyperparameter settings of the best-trained and tested random forest regression and BDTR models are listed in **Table A-3** and **Table A-4**, respectively.

Cross plots that show the mean absolute error against the number of trees constructed by the BDTR model for P_b (complete and incomplete datasets) and for B_{ob} (complete and incomplete datasets) are shown in **Figure A-17**, **Figure A-18**, **Figure A-19,** and **Figure A-20**, respectively. It was observed that by growing the number of trees, the mean absolute error decreases and then converges.

Table 2—Statistical measures for estimating the bubble point pressure for the multiregional dataset.

Correlation	E_a (%)
Labedi	27.82
Kartoatmodjo	28.17
Vasquez and Beggs	32.98
Glasø	34.32
Dokla and Osman	39.82
Standing	25.20
Al-Marhoun	49.61
Frashad	22.73
ANN (Ghrabi *et al.*, 2003)	15.38
Random forest regression	14.13
BDTR (with imputation)	44.06
BDTR	9.17

Table 3—Statistical measures for estimating the oil formation volume factor at the bubble point pressure for the multiregional dataset.

Correlation	E_a (%)
Frashad	2.38
Abdul Majeed	2.53
Kartoatmodjo	2.10
Vasquez and Beggs	3.51
Glasø	4.32
Dokla and Osman	3.21
Standing	2.47
Al-Marhoun	2.35
Petrosky	2.52
Labedi	2.62
Obomanu	2.49
Elsharkawy	3.40
ANN (Ghrabi *et al.*, 2003)	2.04
Random forest regression	1.34
BDTR (with imputation)	3.28
BDTR	0.85

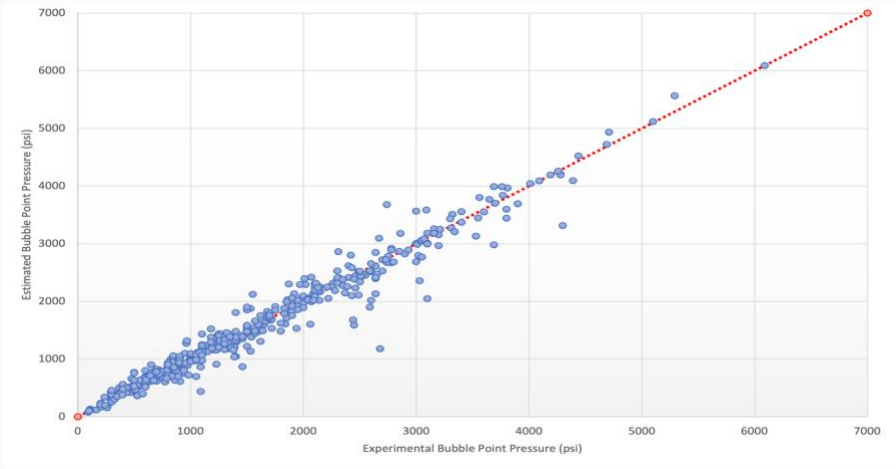

Figure 4—Cross plot of the bubble point pressure for the boosted decision tree regression model (multiregional and complete).

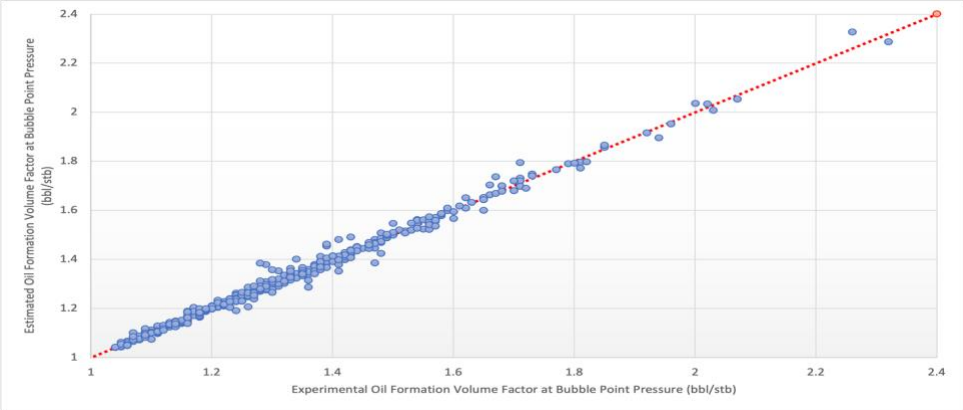

Figure 5—Cross plot of the oil formation volume factor at bubble point pressure for the boosted decision tree regression model (multiregional and complete).

2.7.2. Performance based on the regional dataset

The regional dataset was considered in this study to evaluate the performance of the BDTR model in estimating P_b for crude oil samples collected from Kuwaiti oil wells and compare the performance of the model against the most commonly used empirical correlations used for Kuwaiti oil samples. Empirical correlations usually produce lower errors if applied to regional datasets that have limited ranges of parameters. Our comparative study aimed to show that the BDTR model was accurate even on regional datasets and that it can replace the most commonly used empirical correlations used for that region.

The multiregional dataset was considered in this study to evaluate the performance of the BDTR model as a universal model in estimating P_b and B_{ob} for crude oil samples from all over the world. The BDTR model was retrained and tested on the Kuwaiti crude oil dataset as a regional dataset. The results showed that the BDTR model has a lower E_a compared to the most commonly used empirical correlations for Kuwaiti crude oil systems. The statistical results of the comparative study are shown in **Table 4**. As shown in the table, the BDTR predictive model produced the highest accuracy in estimating P_b compared with the most commonly used

empirical correlations for Kuwaiti crude oil systems. The E_a of the tested correlations and the BDTR model for estimating P_b as a regional dataset are as follows: 6.38 for Lasater's correlation, 6.62 for Frashad's correlation, and 5.65 for the BDTR model.

Cross plots that show the mean absolute error against the number of trees constructed by the BDTR model for P_b (complete and incomplete datasets) are shown in **Figure A-21** and **Figure A-22**, respectively. It was observed that by growing the number of trees, the mean absolute error decreases and then converges.

2.7.3. Performance on incomplete datasets with imputation

PVT reports can include some missing data points, in which case the dataset derived from that report is incomplete. The imputation evaluation was aimed at assessing the performance of the BDTR model when used to estimate the PVT properties of incomplete and imputed datasets.

2.7.3.1. Incomplete multiregional dataset

This case study aimed to show the robustness of the model for diverse datasets with multiple missing data points in the dataset. The results show that estimations of P_b and B_{ob} by the BDTR model trained and tested on an incomplete multiregional dataset in which multiple missing data points were imputed with substitute values have a higher E_a compared to the BDTR model trained and tested on the complete multiregional dataset. As shown in Table 2 and Table 3, the E_a of the tested BDTR model was 44.06 for P_b and 3.28 for B_{ob}

We observed that by applying data imputation to the incomplete dataset of the PVT properties T, Rs, γ_{API}, and γ_g, the built BDTR model remained accurate.

2.7.3.2. Incomplete regional dataset

For the incomplete Kuwaiti crude oil system dataset (the regional dataset), data imputation was applied to 10% of each of the input features with missing values. The data imputation was based on the average value of each of the corresponding features in the entire dataset as well as in the training and testing datasets. The imputed missing data points were included in the test dataset only. The training dataset was complete.

From the input datasets, 59% of the data were used to train and validate the model. The rest of the data were used to test the model as an incomplete dataset with imputation. The training datasets did not include any missing values. The hyperparameter settings are shown in Table A-4.

The results showed that by applying data imputation to the incomplete dataset of the PVT properties T, Rs, γ_{API}, and γ_g, the built BDTR model remained accurate. The built model outperformed the most commonly used empirical correlations in estimating the P_b of Kuwaiti crude oil systems, as in Lasater (1958) and Frashad (1996), in terms of E_a. Lasater's (1958) and Frashad's (1996) correlations are the most commonly used correlations for Kuwaiti crude oil systems at the Kuwait Oil Company.

The statistical results of the comparative study are shown in Table 4 for the incomplete dataset with imputation. As shown in the table, the BDTR predictive model produced the highest accuracy in estimating P_b compared with the most commonly used empirical correlations for Kuwaiti crude oil systems.

Table 4—Statistical measures for estimating P_b for the regional dataset.

Method	E_a (%)
BDTR model (trained on the regional dataset and tested on 48 regional data points with imputation)	25.93
Lasater correlation (tested on 48 data points with imputation)	28.07
Frashad correlation (tested on 48 data points with imputation)	27.06
BDTR model (trained on the regional dataset and tested on 32 regional and complete data points)	5.65
Lasater correlation (tested on 32 regional and complete data points)	6.38
Frashad correlation (tested on 32 regional and complete data points)	6.62
BDTR model (trained on the worldwide dataset and tested on 117 regional data points with imputation—unseen dataset)	18.44
BDTR model (trained on the worldwide dataset and tested on 48 regional data points with imputation—unseen dataset)	26.01

A cross plot of the estimated values of P_b to the corresponding experimental ones is shown in **Figure A-13** for the incomplete dataset with imputation. As shown in this figure, the estimated values by the built BDTR model were in good agreement with the experimental ones.

2.7.4 Training on the multiregional dataset and testing on the unseen and regional dataset

The trained universal model was tested on the Kuwaiti crude oil dataset as a comparative study to showcase the capability of the model in estimating incomplete regional unseen data. A cross plot of the estimated values of P_b to the corresponding experimental ones is shown in **Figure A-14** for the incomplete dataset with imputation. This plot shows the BDTR model that

was trained on the worldwide dataset, being applied and tested on the Kuwaiti dataset as an unseen dataset. As shown in **Figure A-16**, the estimated values by the built BDTR model were in good agreement with the experimental ones.

2.7.5. Feature importance

The PFI score was used to determine the importance of the input features to the estimated values for the regional and multiregional datasets. The tree representation of the model was analyzed in terms of feature importance.

2.7.5.1. Feature importance of the multiregional and regional datasets

The BDTR predictive model for the multiregional dataset determined R_s to be the most important input feature in estimating P_b and B_{ob} (**Table 5**). The PFI score was used to determine the importance of the input features.

As shown in Table 5, T was of great importance to the B_{ob}. However, it was the least important input feature in estimating P_b. As shown in Table 5, γ_{API} ranked as the second most important to P_b but ranked as the third most important feature in estimating B_{ob}. γ_g was of substantial importance for P_b. However, γ_g was the least important input feature for the B_{ob}. R_s was more important in estimating P_b compared with B_{ob}.

As shown in **Table 6**, R_s was determined by the BDTR predictive model for the regional dataset, with imputation as the most important input feature in estimating P_b. We observed that γ_g was of great importance to the P_b, whereas T was ranked as the third most important feature in estimating P_b and γ_{API} was the least important input feature for the P_b.

Tables 5 and 6 show that R_s is the most important feature. The difference in the ranking of the feature importance of the other input parameters was insignificant. This could be due to

the difference in the number of data points and the difference in the range of the values of the input parameters between both datasets, multiregional and regional.

R_s is the solution gas–oil ratio, which indicates how much gas will evolve from the oil under surface conditions. It is a function of T and P_b and greatly affects the specific gravities of gas and oil in a given condition (i.e., T and P_b). Therefore, R_s entails other parameters.

Table 5—Input feature importance to the estimated values for the complete multiregional dataset.

	Input features (PFI scores)			
Output features	R_s	γ_g	T	γ_{API}
P_b	1.4	0.33	0.13	0.47
B_{ob}	1.07	0.09	0.27	0.18

Table 6—Input feature importance to the estimated values for the incomplete regional dataset.

	Input features (PFI scores)			
Output features	R_s	γ_g	T	γ_{API}
P_b	0.25	0.07	0.01	-0.05

2.7.5.2. The tree representation of the boosted decision tree regression model

Based on the model's tree representation, for all the datasets used in this study, we observed that most of the node splits were based on the R_s feature. This was also observed for the root node. Given that the method of splitting the nodes is based on choosing the threshold that would generate the lowest MAE, where each new node or root is chosen if it has the lowest MAE from the node candidates, splitting the nodes based on R_s can reduce MAE in the learning process and thus make more accurate estimations of the target values. This also implies that R_s is the most important feature of the BDTR model. **Figure A-23** shows a tree representation constructed by the algorithm in the gradient-boosting ensemble for B_{ob} using the multiregional and complete dataset, where the root is based on R_s.

2.7.5.3. Solution gas–oil ratio

In the present study, we observed that the higher the value of the P_b or B_{ob}, the higher the relative deviation was between the experimental and the estimated values. Higher values of R_s increase P_b and B_{ob} of different oil samples. A higher value of R_s indicates that a higher amount of gas is dissolved in the oil. At $P > P_b$, the higher dissolved gas could make the models' predictions fluctuate more, resulting in an increased deviation of the predicted results. The data are sparse at high R_s values, which can also cause widening deviations.

2.8. Conclusion

This study has attempted to address two key questions: "Can an accurate BDTR model be built for multi-regional and regional datasets?" and "Which parameter is the most essential in estimating the PVT properties by a BDTR model?" Our findings reveal that the BDTR model consistently outperformed other methods available in the literature in estimating bubble point

pressure and the oil formation volume factor at bubble point pressure. Using multiregional and regional datasets with incomplete parameters, we systematically compared the accuracy of the evaluated model against other methods, such as ANNs, random forest regression, and the most commonly used empirical correlations.

The BDTR model showed superior performance on both multiregional and regional datasets. We showed that the BDTR model's performance was robust, even with missing data points in the dataset. Furthermore, the BDTR model trained on the multiregional dataset was effective and accurate, even when used to estimate unseen and regional datasets. This also showed that the model does not suffer from overfitting. We also analyzed the BDTR model in this study in terms of feature importance based on two approaches: the PFI algorithm and a tree representation analysis. Either approach determined the solution gas–oil ratio to be the most important input parameter for the BDTR model in estimating bubble point pressure and the oil formation volume factor at bubble point pressure. Other studies such as (Gharbi *et al.*, 2003) also found similar trends. Therefore, it is expected to see R_s as the main splitting factor of the nodes in our model. The same trend is observed for the root nodes which yet again indicates the importance of this parameter on the final predicted values.

The proposed BDTR model is preferable due to its interpretable tree representation, which can facilitate the understanding of the key input features. Data analytics in petroleum engineering relies on the incorporation of interpretable algorithms to facilitate predictive models. The results in this study provide a strong foundation for future work on the physical meaning behind the interpretability of the model. This type of model can be extended to predict other petrophysical properties, such as permeability and porosity, for further estimations in upstream exploration and production computations. The model can be retrained and used to study the gas–

liquid multiphase flow regimes, which are currently explored. Lastly, this model can be integrated into the upstream oil and gas simulators used for reservoirs' prediction and monitoring activities.

Chapter 3

Estimation of Liquid Holdup in Gas–Liquid Two-Phase Flow

3.1 Introduction

Early detection and identification of liquid holdup (H_L) in oil and gas wells are needed to reduce the maintenance downtime and thus increase production. This is crucial in designing oil and gas facilities. Consequently, many studies on predicting liquid holdup have been carried out, considering different flow conditions, resulting in a large number of empirical correlations with various degrees of accuracy.

Machine learning approaches in predicting liquid holdup in multiphase flows have recently been studied to improve the prediction accuracy compared to the existing empirical correlations. However, these approaches ignore the heuristic feature importance of the input parameters to the predicted H_L values. The heuristic feature importance can help in gaining insight into the issues associated with the H_L studies, such as the liquid loading phenomenon.

In our study, a machine learning predictive model, BDTR, was trained, tested, and evaluated in predicting H_L in multiphase flows in oil and gas wells. The predictive power of the BDTR model has been studied and verified before in our preliminary studies on PVT datasets (Almashan, Naruse and Morikawa, 2019). In the present study, the development of the predictive

model was based on experimentally derived datasets that were collected from available literature. Air–kerosene and air–water mixtures were used in deriving the 111 experimental data points. Implementing the BDTR predictive model can help in understanding the feature importance by referring to the decision trees constructed by the model as earlier splits based on certain features' values and can thereby infer that these features are in particular more important than the subsequent ones. To the best of our knowledge, the present study is the first work that shows how the decision forest regression predictive models can accurately predict H_L.

Our results show that the proposed BDTR model outperforms the most commonly used empirical correlations and the fuzzy logic model used in estimating H_L in gas–liquid multiphase flows. For the built model, the most important input feature in estimating H_L is the superficial gas velocity (v_{sg}). The proposed model can be extended to predicting other properties or solving other regression problems in the upstream oil and gas processes. In addition, the proposed predictive model can easily be integrated into simulators.

3.2 Literature review

Building offshore oil and gas facilities that include transportation pipelines, multiphase separators, process plants, oil and gas well completion, and tubing equipment requires an accurate estimation of liquid holdup in multiphase flow. In 1958, Flanigan conducted a field study of a 16-in pipeline with measurements of pressure drops of different pipeline sections. He found no evidence of pressure recovery in the downhill sections, and hill inclination did not affect the pressure drop. The panhandle equation was used for the friction loss calculation. The loss caused by elevation was calculated using an elevation factor. In developing the correlation, the overall pressure drop was considered. Therefore, in Flanagan's methods, the elevation term includes any pressure recovery that may have been present in the study.

In 1967, Guzhov, Mamayev and Odishariya presented a study on inclined two-phase flow. The authors acquired the datasets from a 2-in pipeline with an inclination between horizontal and 9 degrees. Two flow patterns were considered as stratified and plug in the development of the correlation. The flow pattern was predicted using the gas input fraction and a mixture Froude number. The authors found that a single liquid holdup expression describes the downhill flow and another the uphill flow, both in the stratified flow.

Payne *et al.* (1979) in 1979 designed and constructed a 550-ft (168-m) long and 2-in (5.1-cm) diameter pipeline in a hilly terrain configuration. Experimental gas-water datasets were obtained and used in evaluating pressure drop and two-phase flow liquid holdup correlations. The authors concluded that the most accurate estimations were obtained by using Beggs and Brill's correlation and by using a combination of Guzhov, Mamayev and Odishariya's and Beggs and Brill's correlations (Payne *et al.* 1979).

Mukherjee *et al.* (1983) in 1983 studied the liquid holdup behavior in inclined two-phase flows with a simulator of an inclined pipeline. Palmer (1975) stated that liquid holdup is overpredicted with the Beggs and Brill correlation in both downhill and uphill flows. Guzhov, Mamayev and Odishariya (1967) developed an inclination-independent liquid holdup correlation. Hughmark and Pressburg (1961) proposed a general gas–liquid multiphase flow liquid holdup correlation that covers a wide range of diameters and physical properties.

Kjolaas *et al.* (2015) conducted investigations on several solutions for liquid holdup occurring in low liquid loading flows. For their study, the authors used a large-scale multiphase flow loop that belongs to an independent research organization in Norway (SINTEF). The diameters of the facility's multiphase flow loop pipelines are 0.2 and 0.3 m. The inclinations were 1°, 2.5°, and 5° degrees in addition to the horizontal orientation. A glycerol–water mixture, water, Exxsol D40, and naphtha were used as the liquid phases. For the gas phase, nitrogen was used. The authors concluded that the accumulation of liquid happens at a constant gas Froude number for a given liquid flow rate and a given pipeline inclination.

One of the multiphase flow correlations that are applicable to inclined, vertical, and horizontal multiphase flow modeling schemes is Beggs and Brill's multiphase flow correlation. Therefore, this correlation is widely used in the industry as a multiphase flow correlation (Sami and Aqqour, 1992). In addition, the different flow patterns of the vertical and horizontal flows are considered by this multiphase flow correlation (Beggs and Brill, 1973). To calculate the pressure gradient, this correlation applies the average in-situ density and the general mechanical energy balance. However, a limitation in the performance of the (Beggs and Brill, 1973) multiphase flow correlation occurs when the parameters of the flow patterns classifications are near the boundaries of the flow patterns. The inaccurate estimation of the slip liquid holdup is the

cause of this limitation (Brill and Mukherjee, 1999). Rammay, Hussain and Al-Nuaim (2015) in 2015 designed and tested a fuzzy logic model to estimate liquid holdup and to predict horizontal flow patterns. In the development of the fuzzy logic model, the learning-by-example technique was used. This is a fuzzy-systems rule-based construction in automated methods (Ross, 2004). In their study, the authors also proposed a modification to the widely used Beggs and Brill multiphase flow correlation.

Wen *et al.* (2017) in 2017 obtained 431 datasets from pipes of 75 mm and 60 mm diameters. In addition, the authors evaluated six of the existing models in estimating the liquid holdup of gas–liquid two-phase flow, using the obtained datasets, in inclined and horizontal pipelines (Wen *et al.*, 2017). Moreover, the model with the best performance was stated in the study. Furthermore, the gas–liquid ratio, pipe diameter, pipe inclination, and liquid types effects on liquid holdup were investigated in Wen *et al.* (2017). The authors concluded that the modified Hughmark correlation has the highest accuracy for estimating liquid holdup of air–water mixture in horizontal and slightly inclined pipelines. On the other hand, for white oil-air flow, the Beggs–Brill model has the highest accuracy compared to the other evaluated models in Wen *et al.* (2017). Wen *et al.* (2017) state that, as long as the gas–liquid ratio is lower than 100 and stable at the same level, the increase of liquid holdup is proportional to the diameter of the pipeline.

Dabirian *et al.* (2018) in 2018 conducted a theoretical and experimental study on the gas–liquid stratified flow by varying fluid properties such as the liquid viscosity and the gas density alongside parameters such as the liquid and gas velocities. To demonstrate the effect of changing the mentioned parameters on liquid holdup, the authors modified the Taitel and Dukler model (1976) with a new closure relationship for interfacial friction factor and acquired new experimental datasets. For their data acquisition, two experimental facilities were used. These

two facilities included a multiphase flow loop at Tulsa University Separation Technology Projects and a well flow loop of the Institute for Energiteknik. In these two facilities, 0.097-m inner-diameter (ID) pipelines were used to construct these two loops. The length of the loops was, however, different. For the acquired experimental datasets, air and SF6 (ρ = 24 kg/m3) were used as the gas phase, and mineral oil (0.033 and 0.12 Pa·s) and water (0.001 Pa·s and 0.005 Pa·s PAC) were used as the liquid phase. To confirm the existence of the stratified flow in the pipeline, the values of the liquid and gas velocities were chosen manually by the authors. The study shows that with a decreasing superficial gas velocity at constant superficial liquid velocity, the liquid holdup increases. In addition, an increase in the gas density with a decrease in the liquid viscosity reduces the liquid holdup. The authors conducted a comparative study using the experimental datasets against the Taitel and Dukler (1976) model with the modified interfacial friction factors definitions. The comparative study conducted by the authors shows that, for the prediction of the liquid holdup, under a wide range of experimental conditions, the studied closure relationships are not appropriate. Therefore, the authors suggested a new closure relationship for the interfacial friction factor. Gas–liquid stratified flow is the simple multiphase flow regime that is commonly observed in horizontal and near-horizontal pipes (Hanyang and Liejin, 2007). The stratified flow is utilized for several applications, including gas–oil transportation in pipelines, heat exchangers, and process engineering (Bhagwat and Ghajar, 2016). In the petroleum industry, the stratified flow is the most favorable multiphase flow regime owing to the fact that it creates minimal pressure drop, erosion, and equipment failure; avoids intermittent flow; and facilitates phase separation (Dabirian *et al.*, 2017; 2018). A number of modeling and experimental studies have been conducted in the past on the stratified flow pattern in order to better characterize and understand the flow using parameters such as pressure drop,

liquid holdup, liquid entrainment, and the interfacial wave. The stratified flow pattern characterization and transitions in inclined pipelines have been studied theoretically and experimentally by researchers such as Shoham (1982), Barnea *et al.* (1982) Shoham and Taitel (1984), Crawford and Weinberger (1985), and Barnea (1987). In recent years, researchers have applied computational fluid dynamics simulations for a better understanding of the stratified flow pattern by characterizing the stratified flow pattern based on studying the gas–liquid interface and the liquid structure.

Table 7 shows the related works for estimating liquid holdup in gas–liquid two-phase flow.

Table 7—Related works for estimating liquid holdup in gas–liquid two-phase flow.

Approach	Pros	Cons	Applicability	Interpretability
Physical mathematical models	Accurate	Computationally complex Designed for certain settings	Universal if tuned	No
Empirical correlations	Less complex compared to the physical models	Lower accuracy compared to the physical models Limited to certain settings	Not Universal	No
Artificial neural networks, fuzzy logic models, other non-tree-based models	Accurate Not complex	Less accurate compared to the physical models	Universal if retrained	No
BDTR (This research)	More accurate Not complex Heuristic feature importance Interpretability	Less accurate compared to the physical models	Universal if retrained	Yes

3.3 Data acquisition

The performance and the predictive power of a machine learning model greatly rely on the quality and the completeness of the dataset used in building the model. The datasets used in training and testing the predictive model are experimental and they were collected from the literature (111 data points). Air–kerosene and air–water mixtures were used in obtaining the 111 experimental datapoints in the Minami and Brill datasets (1987). In our study, this dataset is divided into three different subsets: training, validation, and testing subsets.

The datasets consist of the following properties: liquid holdup (H_L), superficial gas velocity (v_{sg}), superficial liquid velocity (v_{sl}), *pressure*, and temperature (T). The statistical measures of the datasets are shown in **Table 8**. As can be seen in the table, v_{sg} values range from 1.56 ft/sec to 54.43 ft/sec, v_{sl} values range from 0.017 ft/sec to 3.12 ft/sec, *pressure* values range from 43.7 psia to 96.7 psia, T values range from 76 ^{o}F to 118 ^{o}F, and H_L values range from 0.0082 to 0.45. The superficial gas velocity (v_{sg}), the superficial liquid velocity (v_{sl}), the *pressure*, and the temperature (T) are the input features to the BDTR predictive model. The liquid holdup (H_L) is the output feature. Non-linearity exists between the dependent variable and all the independent variables.

$$H_L = f_3(T, Pressure, v_{sg}, v_{sl}). \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots. (5)$$

Table 8—Statistical measures of the datasets.

Property	Min	Max	Mean
v_{sg} (ft/sec)	1.56	54.43	18.88
v_{sl} (ft/sec)	0.017	3.12	0.38
Pressure (psia)	43.7	96.7	68.7
T (^{o}F)	76	118	105
H_L	0.0082	0.45	0.15

3.4 Experimental setup

3.4.1 Liquid holdup comparative study

Early detection and identification of liquid holdup is needed for safety issues and to reduce maintenance downtime and thus increase production. This is needed in designing subsurface and surface oil and gas facilities. Such facilities include multiphase separators, process plants, and offshore pipelines. Consequently, much research on predicting liquid holdup has been done, resulting in a number of empirical correlations with a varying degree of accuracy. One of the most commonly used empirical correlations for predicting liquid holdup in multiphase flows is the Beggs and Brill correlation (Sami and Aqqour, 1992) for its accuracy in predicting liquid holdup in vertical, horizontal, and inclined pipelines. In addition, the Beggs and Brill correlation considers the different flow regimes in vertical and horizontal flows (Beggs and Brill, 1973). However, there are some limitations to empirical correlations in predicting liquid holdup. These limitations can be observed when the parameters of the flow pattern classifications are close to the flow regime boundaries, as the prediction accuracy decreases.

It has been shown in the past that using empirical corrections as predictive approaches in identifying liquid holdup is effective. Therefore, machine learning approaches in predicting liquid holdup in multiphase flows have been recently studied to improve the accuracy level of estimating the liquid holdup. In the present study, a machine learning predictive model was built to predict liquid holdup in multiphase flows of natural gas in gas wells during exploration and production processes. A machine learning algorithm, BDTR, was trained, tested, and evaluated in predicting the liquid holdup. The datasets used in training and testing the predictive model are experimental and were collected from the available literature. In addition, a comparative study

that includes the most commonly used empirical correlations in multiphase flow for estimating liquid holdup, such as the Beggs and Brill correlation, is conducted. Moreover, in the comparative study a fuzzy Logic model (Rammay *et al.*, 2015) is considered.

3.4.2 Boosted decision tree regressions for estimating H_L

A large set of trees is constructed in decision trees by avoiding correlations among them. The average of the decision trees is chosen as the predictive model in order to avoid overfitting. In boosted decision trees, overfitting is avoided based on limiting the subdivision times and the number of data points in each region. The BDTR algorithm constructs a sequence of trees where each tree in the sequence corrects itself by learning from the error of the previous tree. Decision trees are categorized as non-parametric machine learning models. In decision trees, the data structure of a binary tree is traversed until a decision (leaf-node) is reached. In addition, decision trees are able to represent nonlinear decision boundaries. Decision forest regression predictive models have an ensemble of decision trees. Each tree in the predictive model generates a Gaussian distribution prediction. At the end of the training step of the predictive model, the Gaussian distributions generated by all trees in the model are aggregated into a single distribution.

In the present study, a BDTR predictive model was trained, validated, and tested to estimate the H_L value. In the BDTR model, every tree is dependent on the prior tree. By fitting the residual of the preceding trees, the predictive model learns based on the boosting algorithm. The boosting algorithm utilized in the training of the proposed model is the MART gradient-boosting algorithm. In this algorithm, in every step taken in the training phase of the model, a loss function is used to measure the error that is corrected in the following step. The final optimal tree is an aggregation of the constructed trees.

For the presented BDTR model, the input parameters are the superficial gas velocity (v_{sg}), superficial liquid velocity (v_{sl}), *pressure*, and temperature (T), and the output parameter is the estimated H_L. From the dataset, 80% of the data points were used for training, and 20% of the data points were used for validation and testing. For the incomplete dataset, 70% of the data points were used for training, and 30% of the data points were used for validation and testing. The hyperparameter settings of the best trained BDTR models are shown in **Table A-6**.

The setting of the hyperparameters proceeded as follows:

1. Hyperparameter settings started with a random search followed by a grid search (a limited combination).
2. Fine-tuning was applied to the hyperparameters to obtain the best model.
3. Different numbers of learning rates were tested for each model, and the results were compared for their convergence to the best model.
4. To avoid overfitting, the max-number-of-leaves-per-tree hyperparameter was maintained below the threshold:

   TH_max_num_leaves ≤ 60% training dataset.

3.4.3 Feature importance

The feature importance of the input parameters for the trained predictive model is determined by the PFI algorithm. This algorithm determines the importance scores of the input feature variables for a given dataset. By randomly permuting the values of the input features, the model's sensitivity is measured based on which the importance score of the corresponding feature is determined. By choosing an evaluation metric, the deviation of that metric after the

permutation of the values of a certain feature quantifies the contribution or the importance of that particular input feature of the model in the performance of the estimation of the target feature; in this case, H_L. If permuting the value of an input feature results in a significant reduction in the predictive power of the model, then this feature is important in estimating the output feature H_L.

3.4.4 Incomplete dataset

The experimental setup for the incomplete liquid holdup dataset:

- Data imputation is applied to 17% of each of the input features with missing values in the total test dataset.
- The data imputation is based on the average value of each of the corresponding features in the test dataset.
- The imputed missing data points are included in the test dataset only. The training dataset is complete.
- Missing data points are data points that were removed intentionally and randomly from the dataset, covering a wide range of the data values.
- The number of the missing data points is chosen randomly as missing data points occur randomly in real-life scenarios.

3.5 Evaluation methodology

The following evaluation criteria and measure were considered to comprehensively assess the BDTR model when applied to liquid holdup datasets.

3.5.1 Evaluation criteria

Evaluation criteria:

- Evaluation of the complete and incomplete datasets
- Feature importance

3.5.2 Evaluation measure

The following statistical accuracy measure was used to evaluate the predictive power of the built model:

Mean absolute percentage error (MAPE: Ea)

The relative absolute deviation between the experimental values and the estimated ones is determined by this statistical accuracy measure, where a lower value of Ea% indicates higher accuracy.

3.6 Results and discussion

The liquid holdup dataset was considered in this study to evaluate the performance of the BDTR model against the most commonly used empirical correlations and the fuzzy logic model.

Results show that the built BDTR model outperforms the most commonly used empirical correlations in predicting H_L in terms of the mean absolute percentage error (E_a). A lower value indicates a better correlation between the predicted values and the experimental ones. The statistical results of the comparative study are shown in **Table 9**. As shown in the table, the BDTR predictive model produced the highest accuracy in predicting H_L compared to the most commonly used empirical correlations and the fuzzy logic model.

Table 9—Statistical measures for estimating H_L.

Correlation/Model	E_a (%)
BDTR (complete dataset)	11.66
BDTR (incomplete dataset)	57.98
Fuzzy logic model (Rammay *et al.*, 2015)	12.55
Empirical Correlations:	
Hughmark (1961)	72.7
Hughmark (1962)	66.4
Eaton and Brown (1967)	48.1
Beggs and Brill (1971)	26.26
Taitel and Dukler (1976)	56.84
Brill *et al.* (1981)	59
Mukherjee and Brill (1983)	39.1
Minami and Brill (1987) (1)	14.65
Minami and Brill (1987) (2)	15.74
Abdul-Majeed (1995)	13.93

A cross plot of the predicted values of H_L to the corresponding experimental ones is presented in **Figure 6**, showing that the estimated values by the built BDTR model are in good agreement with the experimental ones.

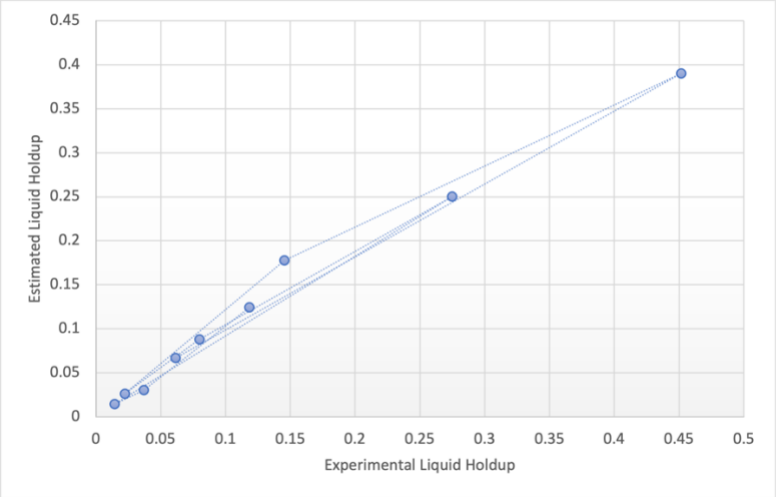

Figure 6—A cross plot of liquid holdup (H_L) for the boosted decision tree regression model.

For the feature importance, the PFI score is used to determine the importance of the input features. As can be seen in **Table 10**, the superficial gas velocity (v_{sg}) is determined by the predictive model as the most important input feature in estimating H_L compared to the other input features used in building this model: the superficial liquid velocity (v_{sl}), *pressure*, and temperature (T).

Table 10—The input feature importance to the estimated values of H_L.

	Input features (PFI scores)			
Output feature	v_{sg} (ft/sec)	v_{sl} (ft/sec)	*Pressure* (psia)	T (^{o}F)
H_L	1.298566	0.28513	0.041088	-0.002804

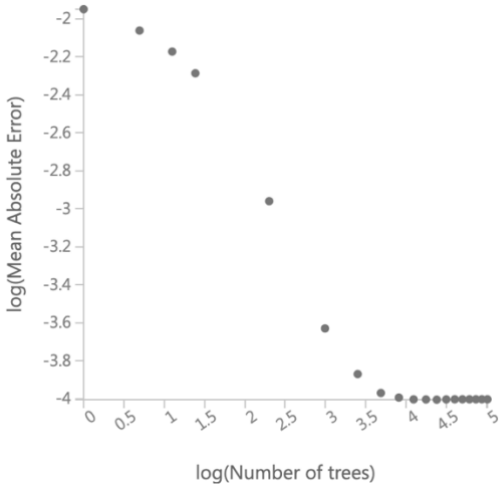

Figure 7—Mean absolute error against number of trees for H_L (complete dataset).

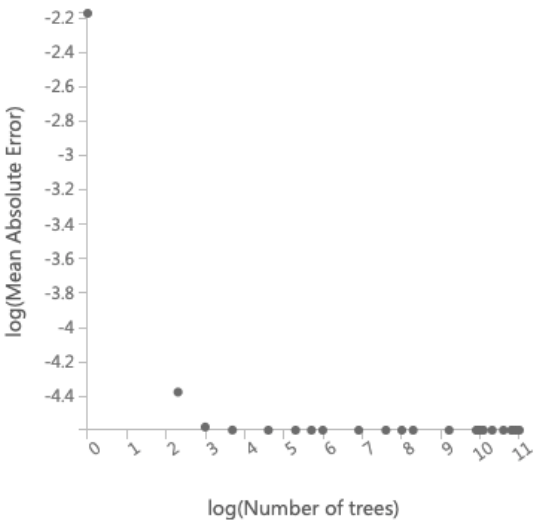

Figure 8—Mean absolute error against number of trees for H_L (incomplete dataset).

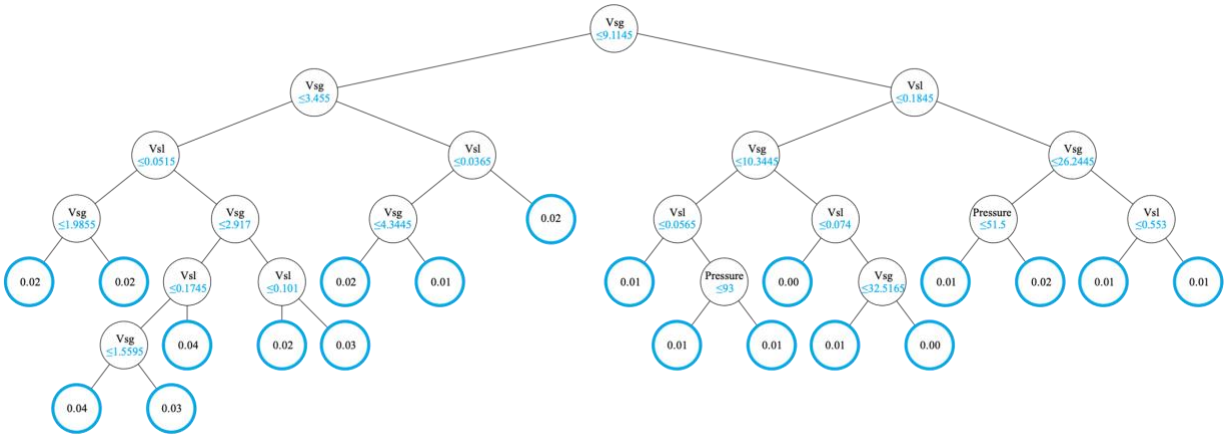

Figure 9—A tree representation from the gradient boosted decision tree ensemble H_L.

3.7 Conclusion

The BDTR predictive model can accurately predict H_L in two-phase flows of gas and liquid in pipelines. In addition, the built BDTR model outperforms the most commonly used empirical correlations and the fuzzy logic model used in predicting H_L in terms of the mean absolute percentage error (E_a).

Furthermore, the built model can be reliably used to obtain heuristic feature importance in an interpretable representation. Moreover, predictions made by the BDTR model can further be investigated by referring to its interpretable representation. For the trained model, the most important feature in estimating H_L is the superficial gas velocity (v_{sg}). As future work, this predictive BDTR model will be studied on the liquid loading phenomenon in gas wells, where different sets of features will be considered. In addition, an extension to the BDTR model can be considered when estimating other features in the upstream exploration and production processes, such as porosity and permeability or the critical flow rate of gas and liquid in production wells. Furthermore, an integration of this model into simulators is possible.

Chapter 4

Gas–Liquid Two-Phase Flow Pattern Classification

4.1 Introduction

During the flow of fluids, a simultaneous flow of two or more phases may occur, called multiphase flow. In the exploration and production of oil and gas wells, multiphase flow causes a deflection in the process of accurately determining the wellbore pressure gradient, which is a critical process in drilling. When the pressure gradient between the wellbore and the reservoir occurs during the kick step of drilling, undesirable fluids and/or gas start to influx into the wellbore, which is known as a multiphase flow regime.

In mechanistic models, the two-phase flow pattern is a key parameter for predicting pressure gradient and liquid holdup, which are both needed to design and operate hydrocarbon production and transportation systems. Every mechanistic model is designed to calculate the pressure drop and liquid holdup per flow pattern. Therefore, the determination of the flow pattern is needed for such models.

In the gas–liquid multiphase flow, the pressure is not unified in the medium due to the difference between the viscosity and density of the liquid and gas. In the case of a vertical flow, the velocity of the liquid and the velocity of the gas are not the same due to the difference in

shear stress. Compared to a single-phase flow, in a multiphase flow, liquid and gas flow separately in smaller areas, which results in a pressure drop for both phases. In contrast to liquids, gas flows with a higher velocity due to its compressibility and smaller density and viscosity. This results in changes in flow patterns from one to another.

In production processes, mixtures of gas and liquids are produced under a large differential in pressure in the pipelines; that is, the difference in pressure between two points along the pipeline. Therefore, accurate determination of the flow pattern and the drop in pressure is crucial to ensure that the design requirements are met, and that production is not affected. This helps improve maintenance and management.

For the determination of the flow patterns, multiphase flow homogeneity—a flow with negligible relative motion between gas and liquid—causes complexity in developing physical equations for determining the flow pattern. Therefore, empirical correlations have been developed. However, the empirical correlations developed in the past for identifying the gas critical flow rate and the multiphase phenomenon can only be applied under certain flow conditions in which they were originally developed; for example, for low or high pressures or for large or small diameters, but not for both conditions.

For any dataset, if the number of examples in each class is not equal, the dataset is imbalanced. Consequently, the distribution of the classes in imbalanced datasets is biased. The bias can be greater if the number of the examples in a certain class is way larger than the other classes in the dataset, or vice versa. Imbalanced datasets can reduce prediction *accuracy* and *precision*, and therefore, the performance of applied machine learning algorithms can significantly be affected. Over- and under-sampling techniques, or resampling techniques, work on adding or removing examples to/from the imbalanced dataset. Over- and under-sampling

techniques help to make the dataset balanced and, as a result, improve the performance of the classification algorithms. The performance of the classification algorithms tends to be better on balanced datasets than on imbalanced ones.

In the present study, the datasets were imbalanced. Therefore, over- and under-sampling techniques were applied to the datasets before training and testing the predictive machine learning models. The datasets used in this study represent 105 data points for low pressure (592 kPa) and 243 data points for high pressure (2,060 kPa), collected from a natural gas facility at Kashiwazaki in Niigata in Japan by the JOGMEC. The facility's pipeline was of large diameter (106.3 mm), and it was inclined at three different angles (0°, 1°, and 3°).

In this study, *k*-nearest neighbors (KNN), decision tree, random forest, and multiclass decision jungle algorithms were trained and tested in gas–liquid two-phase flow pattern classification for imbalanced datasets in which over- and under-sampling techniques were tested and compared. The feature importance of the input parameters is discussed.

The results showed that the best model for classifying the two-phase flow patterns for the low-pressure dataset was the decision tree model with the SMOTEENN preprocessing using over- and under-sampling techniques, with an overall *accuracy* of 91%, a micro-averaged *precision* of 91%, a micro-averaged *recall* of 91%, and an F_1*-score* of 91%. For the high-pressure dataset, the best model for classifying the two-phase flow patterns was the DAGs model with the synthetic minority over-sampling technique (SMOTE) preprocessing over- and under-sampling techniques, which yielded an overall *accuracy* of 91%, a micro-averaged *precision* of 91%, a micro-averaged *recall* of 91%, and an F_1*-score* of 91%.

For the feature importance of the input parameters to the trained DAGs models, v_{sl} was determined to be the most important feature in classifying the flow patterns in the low-pressure

dataset. For the high-pressure dataset, H_L was determined to be the most important feature in classifying flow patterns.

To the best of our knowledge, this is the first study to apply sequential SMOTE, SMOTE, SMOTETomek, and SMOTEENN as over- and under-sampling techniques with KNN and tree-based models for imbalanced datasets for gas–liquid two-phase flow flow-patterns classification in pipelines, in addition to determining the PFI of the input parameters to the DAGs models.

4.2 Literature review

When gas and liquid simultaneously flow in a conduit, the spatial distribution of this phase is referred to as the flow pattern of a multiphase flow. In multiphase flow analysis, flow pattern classification is a fundamental problem, as subsequent processes in the multiphase flow characterization pipeline use the predicted flow pattern in further analysis and applications. In gas–liquid two-phase flow in pipelines, pressure drop determination is crucial for system optimization and pipeline design. Consequently, many researchers have studied gas–liquid two-phase flows in pipelines. They have observed that different flow patterns of gas and liquid occur and change due to different experimental conditions, such as the flow rates of liquid and gas, the temperature, the pressure, the fluid rheology, and the pipe roughness and geometry.

In Beggs and Brill (1973), the authors used test facilities at the University of Tulsa to study two-phase flow in inclined pipelines, focusing on parameters such as flow rates, temperature, pressure, liquid holdup, and different flow patterns that occurred in the conduit. The authors consequently developed correlations to estimate the two-phase flow pressure drop in the pipeline. In Kang and Jepson (2002), the authors conducted a study on multiphase flow with inclined pipelines of a large diameter. In that study, the authors observed that the conditions

under which the flow regime transitions occurred were different in pipelines of large diameters compared to smaller pipelines. In Oddie *et al.* (2003), the authors studied gas-water flow in inclined pipelines with a large diameter, measuring liquid holdup and various flow parameters for a pipeline of 11 meters in length, using 444 tests to observe the different flow regimes of gas–liquid flow.

In Abduvayt *et al.* (2003a; 2003b), the behavior of the two-phase flow in nearly horizontal pipelines under conditions of high pressure was studied. The author further investigated the effects on the flow behavior caused by the diameter of the pipeline and by the pressure. The two-phase flow behavior in vertical pipelines with a large diameter and under high pressures was studied in Omebere-Iyari *et al.* (2007), characterizing and outlining different flow patterns. In Gokcal *et al* (2010), the authors studied oil–gas flows in horizontal pipelines and the effect of oil viscosity. The authors applied a modification to the existing unified model for liquids of high viscosity.

In Khaledi *et al.* (2014), the authors studied two-phase flow patterns in horizontal pipelines and the effect of liquid viscosity. The flow patterns for air and oil with high viscosity two-phase flow were mapped in Vieira *et al.* (2018). The experiments were run using inclined and upward pipelines. A unified model was compared against a commercial dynamic multiphase flow simulator using data that were collected experimentally. The authors stated that the results were in good agreement with the experimental values.

A gas–liquid two-phase flow pattern map was presented in Baker (1954). A two-phase flow pressure distribution estimation model for vertical and inclined pipelines was developed in Gould *et al.* (1974). By using the superficial gas and liquid velocities as coordinates, a horizontal flow pattern map was produced in Mandhane *et al.* (1974).

In Taitel and Dukler (1976), the authors presented a flow pattern map for estimating the transition of flow patterns in two-phase flow horizontal and near-horizontal pipelines. In that study, the coordinates of a two-dimensional plot were calculated using experimental data. Subsequently, the boundaries of the flow patterns were determined on the map. The flow patterns considered in that study included stratified wavy, stratified smooth, intermittent, annular-annular dispersed, and dispersed bubble flow patterns.

In Arya *et al.* (1981), two-phase pressure drop and liquid holdup correlations were compared against one another across the boundaries of the flow patterns. The authors concluded that discontinuities existed across the boundaries of the flow patterns.

In Barnea *et al.* (1982), the authors developed inclined pipeline flow pattern transition estimation mechanistic models. The authors in Brito *et al.* (2015) presented a flow pattern classification model for horizontal pipe gas–viscous-liquid flows. In that study, three different gas–liquid experiments were conducted. The authors concluded that, based on a comparative study, the proposed model outperformed existing mechanistic models for systems of gas and liquid with high viscosity.

Flow pattern transition prediction models in horizontal pipelines for gas–liquid flows of high viscosity were proposed in Al-Safran and Al-Qenae (2018). The authors concluded that the accuracy of their models was higher when the viscosity in the flow was high. The conventional methods used in estimating flow patterns include correlations and semi-analytical methods.

Over the last few years, predictive models based on machine learning algorithms have been studied in estimating the flow patterns in gas–liquid flows. In Li *et al.* (2014), the authors built a model for estimating the gas–liquid two-phase flow patterns in pipelines based on a support vector machine (SVM) supervised learning algorithm. In Mask, Wu and Ling (2019), a

machine learning model based on more than 8,000 multiphase flow laboratory tests was presented. Some of their datasets dated back to the 1950s. The parameters included in the datasets were velocity, pressure, viscosity, density, temperature, surface tension, pipe angle and diameter, and experimentally determined flow patterns. The observed flow patterns included annular, stratified wavy, stratified smooth, dispersed bubble, bubble, and intermittent flows. In that study, the authors stated that the flow pattern was affected by the conduit dimensions, the mechanical properties, the fluid properties, and the in-situ flow rates of gas and liquid. Due to the large difference between the actual flow conditions and the laboratory analysis, the authors derived some dimensionless variables as an approach to characterizing the obtained data points. Based on these dimensionless variables, machine learning flow pattern predictive models were developed. The authors stated that the accuracy of the models increased by more than 90% when the derived dimensionless variables were used. Furthermore, their proposed model outperformed the empirical correlations and semi-analytical models.

These studies did not consider the effect of imbalanced datasets on the performance of the machine learning classification algorithms for classifying the flow patterns. However, in He *et al.* (2021), a two-phase flow pattern identification model of a centrifugal pump based on the SMOTE algorithm and ANNs was used to investigate the flow characteristics of a centrifugal pump under gas–liquid two-phase conditions for imbalanced datasets. In their approach, the authors did not consider under-sampling techniques, and they did not use KNN or tree-based models for classification, as opposed to our approach. In addition, their work is for centrifugal pumps and our work is for oil and gas pipelines, and they studied four flow patterns, whereas in our approach, we are studying eight flow patterns.

To the best of our knowledge, this is the first study to consider evaluating the effect of over- and under-sampling techniques for KNN and tree-based models on gas–liquid two-phase flow-pattern classification in pipelines.

Table 11 shows the related works for gas-liquid two-phase flow classification.

Table 11—Related works for gas–liquid two-phase flow classification.

Approach	Pros	Cons	Applicability	Interpretability
Mechanistic models	Accurate	Computationally complex Designed for certain settings	Universal if tuned	No
k-nearest neighbors, artificial neural networks, other non-tree-based models	Accurate Not complex	Less accurate compared to the mechanistic models	Universal if retrained	No
• Multiclass decision jungle • Decision tree • Random forest	Accurate Not complex Heuristic feature importance Interpretability	Less accurate compared to the mechanistic models	Universal if retrained	Yes

4.2.1 Gas–liquid two-phase flow

The characterization of the flow patterns is based on the relative position of the interface between the liquid phase and the gas phase (Arihara, 2002; Abduvayt, 2003). The flow patterns in this study are defined as follows:

At the dispersed bubble flow, bubbles are created due to liquid turbulence at high liquid flow rates. At this flow, the liquid phase carries many small bubbles, where the two phases—gas and liquid—have no relation in motion. With a pipe of a large diameter (106.3 mm), it has been observed that the number of bubbles increases at high gas flow rates or at low liquid flow rates (Arihara, 2002; Abduvayt, 2003). Previous studies have also demonstrated that the bubble's shape is of an irregular shape and size, being transformed from a regular round shape (Arihara, 2002; Abduvayt, 2003). Consequently, a slip is caused between the two phases. Under high pressure, the distribution of the bubbles is less than that of a lower pressure situation (Arihara, 2002; Abduvayt, 2003).

At a liquid flow rate above 3.00 m/s, the dispersed bubble flow rate can be observed, regardless of the pipe sizes and the pressure rates. The inclination angles of the pipe (0°, 1°, and 3°) do not affect the transition of the dispersed bubble flow. However, a large transition region between the dispersed bubble flow and the intermittent flow can be observed at high-pressure rates (Arihara, 2002; Abduvayt, 2003).

At low liquid and gas flow rates, a stratified flow pattern can be observed. In this flow pattern, gravity separates the two phases. This flow pattern is sub-classified into roll-wave flow, stratified-wave flow, and stratified-smooth flow. The interface between the gas phase and the liquid phase is smooth in a stratified-smooth flow pattern. At a higher gas flow rate, a stratified-

wave flow pattern can be observed. In this flow pattern, waves are formed around the interface between the two phases. In the roll-wave flow pattern, liquid moves up the wall of the pipe, and concave-down curvatures can form around the interface of the gas phase and the liquid phase (Arihara, 2002; Abduvayt, 2003).

In pipelines of small diameters and at low pressure rates, a stratified-flow pattern can be observed, in which the level of the liquid phase is lower than the centerline of the pipe. However, for pipelines of small diameters and at low pressure rates and when the superficial liquid velocity is above 0.08 m/s, this stratified flow pattern may not be observed. Based on the liquid holdup datasets, some data points of high-pressure rates reflect the existence of a stratified flow pattern with a liquid level that is above the pipe centerline (Arihara, 2002; Abduvayt, 2003).

Pipelines of large diameters can have a stratified flow pattern when the superficial liquid velocity is at 0.0156 m/s. Furthermore, the stratified flow pattern can be greatly affected by the pipeline angle of inclination, in which higher angles can cause liquid holdups due to the slower motion of liquid, which will prevent the stratified flow pattern from occurring inside the pipeline. However, with pipelines of large diameters at an inclination angle of 1° and at a low rate of pressure, a stratified roll-wave flow pattern can be observed when the superficial liquid velocity is low and the superficial gas velocity is high (Arihara, 2002; Abduvayt, 2003).

For pipelines of an inclination at 1° and with a high rate of pressure in the transition region between the stratified flow pattern and the annular flow pattern, flow patterns that are a mixture of stratified and annular flow patterns can occur. However, as mentioned earlier, under similar conditions with an inclination angle of 3°, no stratified flow pattern occurs (Arihara, 2002; Abduvayt, 2003).

An alternating flow pattern between liquid and gas is defined as an intermittent flow pattern. It is sub-classified as an elongated bubble flow pattern, slug flow pattern, and froth flow pattern. In the elongated bubble flow pattern, gas pockets exist in the liquid slugs that occupy the entire pipe cross-section. In this flow pattern, liquid slugs either contain very few bubbles of gas or contain no bubbles at all. It has also been observed that in this flow pattern, gas pockets tend to exist at the upper half of the channel of flow (Arihara, 2002; Abduvayt, 2003). When the pressure rate is high and the pipe diameter is large, the elongated bubble flow pattern occurs at high superficial liquid velocities but not in low-pressure-rate and small-diameter cases (Arihara, 2002; Abduvayt, 2003).

One of the main characteristics of the slug flow pattern is the alternating flow of the gas and liquid phases. A slug unit consists of a body and a film. The slug body may contain small bubbles of gas that are more toward the ceiling of the pipeline or toward the front side of the body of the slug (Arihara, 2002; Abduvayt, 2003).

The intermediate flow pattern between the annular flow pattern and the slug flow pattern is the froth flow pattern. The froth flow pattern in this study is not related to the froth flow pattern in vertical or near-vertical pipelines, or to the churn flow pattern. It is, instead, a flow pattern that is very similar to the slug flow pattern mentioned earlier but with a higher gas rate that causes the slugs of liquid to split in the plugs of gas. Consequently, the plugs of gas change into irregular and smaller shapes. Further, the slugs of liquid become broken by regions that are highly concentrated with gas (Arihara, 2002; Abduvayt, 2003).

Previous studies have shown that for each pipeline inclination angle (0°, 1°, and 3°), and under a high rate of pressure, the froth flow pattern and the slug flow pattern can be observed at

low superficial gas velocities, as opposed to the low rate of pressure case (Arihara, 2002; Abduvayt, 2003).

The annular flow pattern occurs at high gas flow rates. This flow pattern has been observed in pipelines of a large diameter and under a low rate of pressure. When this flow pattern has been visually observed, the majority of the liquid has seemed to flow in the lower half of the pipeline as a film accompanied by small waves flowing around the inner section of the pipeline flow path. To some extent, the annular flow pattern is similar to the stratified flow pattern, as the majority of liquid flows at the lower half of the pipeline, with a small portion flowing as waves touching the wall of the pipeline and forming films of liquid. This flow pattern with higher liquid flow rates makes it similar to the froth flow pattern (Arihara, 2002; Abduvayt, 2003).

An accurate estimation of liquid holdup and pressure drop can be achieved by identifying the flow patterns due to their dependence on each other. In horizontal or near-horizontal pipelines, the distribution of the film thickness is not uniform, as opposed to vertical or near-vertical pipelines (Arihara, 2002; Abduvayt, 2003). The flow patterns in the present study were observed at angles of 0°, 1°, and 3°, and they were defined by Abduvayt (2003).

4.3 Methodology

4.3.1 Multiclass decision jungle

A decision jungle consists of an ensemble of decision DAGs. They are considered an extension of random forests. In a decision jungle, multiple paths from the root to each leaf are allowed by a DAG, as opposed to conventional decision trees, in which only a single path is allowed to every node (Shotton *et al.*, 2013).

The decision jungle's advantages are as follows:

- A decision DAGs model has better generalization performance and a lower memory footprint compared to a decision tree. This is due to the merged tree branches, which are allowed only by a decision DAGs model.
- Nonlinear decision boundaries can be represented by non-parametric decision jungles.
- Decision jungles are noisy features resilient. The classification and feature selection are integrated into the model's performance.

4.3.2 Random Forest algorithm

Various tree-based algorithms can be used, such as decision trees, random forest, bootstrap aggregation (bagging), and gradient boosting. Random forest is an efficient tree-based supervised learning algorithm that is commonly used. It is one of the best algorithms because it is simple to use. The algorithm has the capacity to perform both classification and regression. Random forest combines hundreds of decision trees and then trains each decision tree on a subset of data from the entire dataset. The final predictions are determined by averaging over a

collection of decision trees. The advantages of random forests are multiple. The independent decision trees appear to overfit the training data but averaging the predictions from different trees by random forest mitigates this overfitting problem. This process yields greater predictive precision than using an individual decision tree to make predictions. Using the random forest algorithm can yield the feature importance of the datasets. This is possible by implementing the Boruta algorithm, which can determine the feature importance of a provided dataset. The random forest algorithm works as follows:

Random samples are picked from the given dataset. For each selected sample, a decision tree is built by the algorithm. A prediction is the outcome of each generated decision tree. Thereafter, voting is carried out for each prediction. The mean is used for regression problems, while the mode is used for classification problems. At the end, the final prediction is determined by the algorithm as the most voted.

4.3.3 Over- and under-sampling

In imbalanced datasets, the number of examples in the training dataset is not equal for each class. Consequently, the distribution of the classes is skewed or biased. This can significantly affect the performance of applied machine learning algorithms, as imbalanced datasets can reduce prediction *accuracy* and *precision*.

The class distribution of a dataset can be changed by removing or adding examples from/to the imbalanced dataset by applying resampling techniques. After generating a balanced dataset, the applied machine learning algorithms' prediction *accuracy* and *precision* can increase, as the performance of the machine learning algorithms tends to be better on balanced datasets than on imbalanced ones. Under-sampling techniques are used with majority classes where examples are merged or deleted. By contrast, in over-sampling techniques, new synthetic

examples are created and added to the minority class. The dataset can be more balanced if both techniques are used together.

The SMOTE algorithm is one of the most frequently used algorithms in resampling imbalanced datasets. The algorithm picks KNN and applies it to create synthetic samples in space based on the feature vectors. Another example is the SMOTE + Tomek links (SMOTETomek) resampling method. A combination of SMOTE and the edited nearest neighbors (ENN) under-sampling method (SMOTEENN) is another effective resampling method.

By applying the SMOTE method, noisy samples may be generated by inserting new points between marginal inliers and outliers. To solve this issue, after implementing SMOTE for over-sampling, the generated space must be cleaned. This can be achieved by applying Tomek links (in SMOTETomek) or ENN (in SMOTEENN) as under-sampling cleaning methods. Establishing well-defined clusters can improve classification performance.

In SMOTETomek, Tomek links are used as a cleaning under-sampling method after applying the over-sampling method SMOTE. The examples in the majority class that form a Tomek link are removed, as one of them is considered noise or an overlapping example. If a Tomek link is formed by two examples in two different majority classes, both examples are removed because they are both near the border.

In SMOTEENN, $k = 3$ nearest neighbors is used to find, in a given dataset, misclassified examples by at least two of the three nearest neighbors, and these are consequently removed from the dataset. This can be applied to both majority and minority classes. Further, in the majority class, ENN applies the KNN method, and if the prediction is different from the majority class, that instance is removed from the dataset. ENN is more aggressive in under-sampling

compared to the Tomek links, as it tends to remove more examples from the space. This can ensure deeper data cleaning.

4.3.4 Permutation feature importance

The feature importance of the input parameters for the trained predictive model was determined by the PFI algorithm. This algorithm determines the importance scores of the input feature variables for a given dataset. By randomly permuting the values of the input features, the model's sensitivity is measured based on which the importance score of the corresponding feature is determined. By choosing an evaluation metric, the deviation of that metric, after the permutation of the values of a certain feature, quantifies the contribution or the importance of that particular input feature of the model in the performance of the estimation of the target feature, which in this case is the flow pattern. Therefore, if permuting the values of an input feature results in a significant reduction in the accuracy of the model, then this input feature is important in estimating the target feature.

4.4 Data acquisition

4.4.1 Test facility

The main objective of the experimental program was to collect data on gas–liquid two-phase flow at low and high pressures in slightly inclined, large-diameter pipes. For this reason, the JOGMEC's large-scale multiphase flow-test facility at Kashiwazaki in Japan was used (Arihara, 2002; Abduvayt, 2003).

To study multiphase fluid flow under close-to-actual well and pipeline flow conditions and to validate multiphase flow development equipment, a multiphase flow experimental facility

was built (Abduvayt, 2003). They used experimental data from this facility to refine the multiphase fluid flow models used in simulations. The facility had a flow loop of 500 m in length with a diameter of 4.19 in and a flow loop of 200 m in length with a diameter of 2 in (Abduvayt, 2003).

The following test sections were included in the facility:

- Test section 1 (inclined): A subsea or onshore hilly terrain pipeline simulation
- Multiphase equipment test section (e.g., multiphase meters and multiphase pumps)
- Test section 2 (inclined): An offshore platform's riser pipe or well simulation.

4.4.2 Supply system

Nitrogen and water were used as gas and liquid phase test fluids, respectively. The gas phase was supplied by a two-stage compressor to the 152.4 mm (6 in) gas supply line—one for high pressure and the other for low pressure—and the water phase contained in the 27.5 m^3 steel tank was supplied by a pump to the 203.2 mm (8 in) water supply line. Gas and liquid were combined in the mixing tee, and the mixture was pumped into the 106.3 mm (4.19 in) ID and 100 m total length test lines (Abduvayt, 2003).

4.4.3 Test matrix

For the low-pressure (592 kPa) and high-pressure (2,060 kPa) conditions, 35 and 81 runs were carried out, respectively. The superficial gas velocity (v_{sg}) ranged from 2.0 m/s to 11.0 m/s for low-pressure conditions, while the superficial liquid velocity (v_{sl}) ranged from 0.16 m/s to 5.7 m/s. v_{sg} ranged from 0.1 m/s to 3.2 m/s for high-pressure conditions, while v_{sl} ranged from 0.04 m/s to 5.6 m/s. The effect of the small inclination angle on the transition boundaries of the flow

pattern was examined by varying the angle of inclination as 0◦, 1◦, and 3◦. The temperature of the fluid was maintained at 35 ± 5 ◦C. **Table 12** shows the test matrix (Abduvayt, 2003).

Table 12—Test matrix.

Test	Diameter (in.)	*T* (°C)	*Pressure* (kPa)	θ (degree)	v_{sl} (m/s)	v_{sg} (m/s)	Data-points
High pressure	4.19 (106.3 mm)	35 ± 5	2,060	0°	0.04 ~5.6	0.1 ~3.2	81
				1°			81
				3°			81
Low pressure			592	0°	0.16 ~5.7	2.0 ~11.0	35
				1°			35
				3°			35

4.4.4 Test procedure

As an experimental technique, the pressure in pipes was first equalized in an experimental condition after commencing the experiment method, and data were acquired with no flow in the pipes. Such data were used for the error analysis of meter measurements. Next, two-phase flow tests were carried out. The gas and liquid flow rates were achieved at the desired values by changing the gas and liquid valves via an automatic control system. If flow was stabilized, data acquisition as commenced. In each experiment, a liquid sample was gathered, and the conductance and temperature were measured. In the two-phase flow experiments for each inclined angle, a single-phase liquid flow experiment was also performed. These data were used to predict the roughness of the pipe wall and the calibration of sensors for liquid holdup (Abduvayt, 2003).

4.4.5 Data analysis

TEXELATM's computer data-processing framework was used to generate experimental datasets with unique features from which flow structure information was extracted. The

specification of the flow patterns for the pipe flow at low pressures can typically be rendered by visual observation of the distribution of gas and liquid molecules within the pipe and by using the holdup sensor signals (Abduvayt, 2003). However, the flow cannot be visually observed under high-pressure conditions (Abduvayt, 2003). Based on changes in the void fraction or the liquid holdup inside the cross section of the pipe, the flow patterns under high-pressure conditions were determined (Abduvayt, 2003). For individual flow patterns, liquid holdup signals have special common features at low and high pressures (Abduvayt, 2003). The flow patterns that were determined in the experiments are described in Section 4.2.1.

4.4.6 Datasets for training machine learning models

In this study, machine-learning-based models were trained, tested, and evaluated to predict gas–liquid two-phase flow patterns. The datasets used in this study represent 105 data points for low pressure (592 kPa) and 243 data points for high pressure (2,060 kPa), which were experimentally collected from a natural gas facility at Kashiwazaki in Niigata, Japan, by the JOGMEC. A total of 3,800 to 5,000 data points were acquired at a sampling rate of 50 to 120 data points per second for each case; 105 and 243 data points are the averages of the total data acquired.

The experimentally derived datasets used in this study include flow pattern, liquid holdup (H_L), temperature, *pressure*, superficial liquid velocity (v_{sl}), and superficial gas velocity (v_{sg}). The statistical measures of the low-pressure (592 kPa) and high-pressure (2,060 kPa) datasets before applying the over- and under-sampling techniques are shown in **Tables 13** and **14**, respectively.

Table 13—The statistical measures of the low-pressure (592 kPa) datasets before applying the over- and under-sampling techniques.

Property	Min	Max	Mean
v_{sg} (m/s)	1.982549	11.025202	5.375185
v_{sl} (m/s)	0.156155	5.669921	1.887716
Pressure (kgf/cm^2)	4.955	7.707	5.192229
T (°C)	10.064	40.283	31.356114
H_L %	8.626	82.332	36.952095

Table 14—The statistical measures of the high-pressure (2,060 kPa) datasets before applying the over- and under-sampling techniques.

Property	Min	Max	Mean
v_{sg} (m/s)	0.0995	3.192066	0.991967
v_{sl} (m/s)	0.043694	5.640249	1.465542
Pressure (kgf/cm^2)	19.097	20.868	19.987417
T (°C)	22.028	46.167	36.102595
H_L %	7.715	99.899	65.216719

For training the predictive models, the number of the datapoints for each angle in the low- and high-pressure datasets are shown in **Table 15**.

Table 15—The number of datapoints of each angle for the low- and high-pressure datasets.

Entire datasets	Angle 0°	Angle 1°	Angle 3°
Low pressure 592 kPa	35	35	35
High-pressure 2,060 kPa	81	81	81

4.4.7 Flow patterns

Flow patterns change based on the gas and liquid flow rates, the properties of fluid and gas, the diameter and incline of the pipeline, and other fluid mechanical properties (Abduvayt, 2003).

As shown in **Figure 10**, the flow patterns in this study are classified as follows: annular (A); dispersed bubble (DB); intermittent (I)—subdivided into froth (F), slug (SL), and elongated

bubble (EB); and stratified (S), subdivided into roll-wave (SR), stratified-wave (SW), and stratified-smooth (SS) (Abduvayt, 2003).

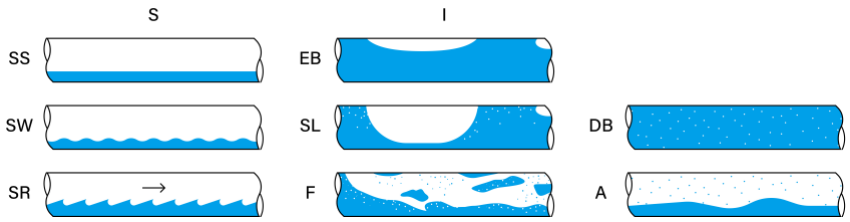

Figure 10—Flow patterns.

Other flow patterns have been observed that are considered a combination of two basic flow patterns, such as intermittent-dispersed bubble flow (In-DB), stratified-wave–stratified roll-wave flow (SW-SR), and stratified roll-wave annular flow (SR-A) (Abduvayt, 2003).

4.5 Experimental setup

The experimental work in this study was divided into two main steps for each dataset (the high-pressure dataset and the low-pressure dataset). The two steps were the data-preprocessing step and the modeling step.

In the data-preprocessing step, the imbalanced datasets were oversampled/over-under-sampled using four different approaches. The first is a sequential SMOTE technique, in which each imbalanced dataset was oversampled using SMOTE, one class at a time, in a sequential manner. In some cases, to improve classification *accuracy* and *precision*, some classes were under-sampled. The second preprocessing approach applied SMOTE to over-sample all the minority classes in the dataset altogether at one time. The third preprocessing approach applied SMOTETomek for over- and under-sampling all the classes in the dataset altogether at one time. The fourth preprocessing approach applied SMOTEENN for over- and under-sampling all the

classes in the dataset altogether at one time. The total number of the datapoints in each dataset before and after applying the over- and under-samplings techniques are shown in **Tables 16–19**.

Table 16—The total number of data points in each dataset before and after applying the sequential SMOTE.

Training datasets	Before SMOTE (sequential)	After SMOTE (sequential)
Low pressure 592 kPa (0°, 1°, and 3°)	73	140
High pressure 2,060 kPa (0°, 1°, and 3°)	170	260

Table 17—The total number of the datapoints in each dataset before and after applying SMOTE.

Training datasets	Before SMOTE	After SMOTE
Low pressure 592 kPa (0°, 1°, and 3°)	73	210
High pressure 2,060 kPa (0°, 1°, and 3°)	170	568

Table 18—The total number of the datapoints in each dataset before and after applying SMOTETomek.

Training datasets	Before SMOTETomek	After SMOTETomek
Low pressure 592 kPa (0°, 1°, and 3°)	73	210
High pressure 2,060 kPa (0°, 1°, and 3°)	170	564

Table 19—The total number of the datapoints in each dataset before and after applying SMOTEENN.

Training datasets	Before SMOTEENN	After SMOTEENN
Low pressure 592 kPa (0°, 1°, and 3°)	73	182
High pressure 2,060 kPa (0°, 1°, and 3°)	170	482

The datasets were split as 70% for training and 30% for validation and testing in DAGs models. In KNN, decision tree, and random forest models, the datasets were split as 70% for training and 30% for testing.

In the DAGs models, individual trees were created with the "bootstrap aggregating (bagging) with replacement" resampling method. This is a machine learning ensemble meta-algorithm that can help improve the accuracy of the machine learning algorithms implemented in statistical regression and classification. Bagging can also help avoid overfitting and reduce variance. A new set of samples was used to build each tree using this method. This subset of samples was chosen based on sampling the main dataset in a random way, with replacement. Duplicates may occur in any subset due to resampling with replacement technique. The sampling with replacement step continued until the size of the new subset was the size of the original dataset that was being sampled. Each subset was independent of its peers. Lastly, aggregation was implemented by combining the generated models based on voting.

The training-testing splits of the datasets were implemented manually to ensure that all the classes (flow patterns) were included in both the training and the testing splits, and that there was no missing class in any of the splits.

The hyperparameters of the DAGs model are as follows:

- Number of decision DAGs: for the ensemble, specify the maximum number of graphs that can be generated.
- Maximum depth of the decision DAGs: each graph's maximum depth should be defined with this hyperparameter.
- Maximum width of the decision DAGs: each graph's maximum width should be defined with this hyperparameter.
- Number of optimization steps per decision DAG layer: when creating each DAG, specify how many iterations over the data should be performed.

The hyperparameter settings for the best-trained DAGs models on the datasets generated by the over- and under-sampling techniques are shown in **Tables 20–23**.

Table 20—The hyperparameter settings for the best-trained models on the datasets generated by the sequential SMOTE.

Datasets	No. of optimization steps per decision DAG layer	Max width of the decision DAGs	Max depth of the decision DAGs	Number of decision DAGs	Resampling method
Low–pressure 592 kPa (0°, 1°, and 3°)	4,595	7,004	407	963	Bagging
High–pressure 2,060 kPa (0°, 1°, and 3°)	5,576	4,600	397	1,311	Bagging

Table 21—The hyperparameter settings for the best-trained models on the datasets generated by SMOTE.

Datasets	No. of optimization steps per decision DAG layer	Max width of the decision DAGs	Max depth of the decision DAGs	Number of decision DAGs	Resampling method
Low–pressure 592 kPa (0°, 1°, and 3°)	7,216	4,470	619	258	Bagging
High–pressure 2,060 kPa (0°, 1°, and 3°)	4,748	4,809	449	185	Bagging

Table 22—The hyperparameter settings for the best-trained models on the datasets generated by SMOTETomek.

Datasets	No. of optimization steps per decision DAG layer	Max width of the decision DAGs	Max depth of the decision DAGs	Number of decision DAGs	Resampling method
Low–pressure 592 kPa (0°, 1°, and 3°)	2,327	3,598	414	155	Bagging
High–pressure 2,060 kPa (0°, 1°, and 3°)	5,037	5,100	433	199	Bagging

Table 23—The hyperparameter settings for the best-trained models on the datasets generated by SMOTEENN.

Datasets	No. of optimization steps per decision DAG Layer	Max width of the decision DAGs	Max depth of the decision DAGs	Number of decision DAGs	Resampling method
Low–pressure 592 kPa (0°, 1°, and 3°)	4,947	4,567	502	180	Bagging
High–pressure 2,060 kPa (0°, 1°, and 3°)	5,199	4,906	496	193	Bagging

4.6 Evaluation measure

The evaluation metrics used in this study for evaluating the classification models were *accuracy*, *precision*, *recall* and *F_1-score*. For imbalanced datasets, as in this study, *recall* was the main metric considered.

The number of positive class predictions that belonged to the positive class was quantified by *precision*. The number of positive class predictions made out of all positive examples in the dataset was quantified by *recall*. The *F_1-score* combines the comments that can be inferred by the quantifications made by both *recall* and *precision* into a single score.

Accuracy is calculated by dividing the number of correct predictions by the total number of predictions, as shown in Equation (6).

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (6)$$

Precision is a metric that quantifies the number of positive predictions made correctly. Therefore, *precision* measures the *accuracy* of the minority class. It is determined as the ratio of positive examples correctly predicted divided by the total number of positive examples that have been predicted. *Precision* assesses the fraction of correct instances classified as positive among those classified.

The micro-averaged *precision* is calculated in imbalanced multiclass problems by summing true positives across all classes and then dividing the summation by the sum of true positives and false positives across all classes, as shown in Equation (7).

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

The micro-averaged *recall* is a statistical evaluation metric that can be used in multiclass classification problems. Out of all the positive predictions made, *recall* quantifies the number of accurate positive predictions made. *Recall* offers an indication of missing positive predictions, unlike *precision*, which focuses only on the right positive predictions out of all positive predictions. Thus, *recall* provides some concept of the positive class's coverage. Typically, *recall* is used for imbalanced learning to assess the coverage of the minority class.

In imbalanced multiclass problems, *recall* is calculated by summing true positives across all classes, and then diving the summation by the sum of true positives and false negatives across all classes, as shown in Equation (8).

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (8)$$

The *F_1-score* is the harmonic mean of *precision* and *recall*. It is considered one of the most commonly used metrics for evaluation in problems with imbalanced classifications. Both scores, *precision* and *recall* are combined in calculating the *F_1-score*, as shown in Equation (9).

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (9)$$

The contributions of all classes are aggregated if micro-average is used to compute the average metric of *precision*, *recall*, or *F_1-score* in multiclass problems. Therefore, in imbalanced multiclass classification problems, the micro-average is preferred.

4.7 Results and discussion

The results showed that the best model for classifying the two-phase flow patterns for the low-pressure dataset was the decision tree model with the SMOTEENN preprocessing over- and under-sampling techniques, which yielded an overall *accuracy* of 91%, a micro-averaged *precision* of 91%, a micro-averaged *recall* of 91%, and an *F_1-score* of 91%, as shown in **Tables 24–31**. The second-best model for classifying the two-phase flow patterns for the low-pressure dataset was the DAGs model with sequential SMOTE preprocessing over- and under-sampling techniques. The model yielded an overall *accuracy* of 86%, a micro-averaged *precision* of 86%, a micro-averaged *recall* of 86%, and an *F_1-score* of 86%, as shown in Tables 24–31. For the high-pressure dataset, the results showed that the best model for classifying the two-phase flow patterns was the DAGs model with the SMOTE preprocessing over- and under-sampling techniques, which yielded an overall *accuracy* of 91%, a micro-averaged *precision* of 91%, a micro-averaged *recall* of 91%, and an *F_1-score* of 91%, as shown in Tables 24–31. Lastly, the second-best model for classifying the two-phase flow patterns for the high-pressure dataset was

the DAGs model with the SMOTETomek preprocessing over- and under-sampling techniques, yielding an overall *accuracy* of 90%, a micro-averaged *precision* of 90%, a micro-averaged *recall* of 90%, and an F_1*-score* of 90%, as shown in Tables 24–31.

The overall *accuracy* of the trained models for the low-pressure (592 kPa) and high-pressure (2,060 kPa) datasets are shown in Tables 24 and 25, respectively.

As shown in Table 24, for the low-pressure dataset, sequential SMOTE, SMOTE, SMOTETomek, and SMOTEENN resampling techniques helped in improving the *accuracy* of the DAGs model. However, for the KNN model, resampling did not improve the *accuracy* of the model. For the Decision Tree (DT) model, only SMOTEENN helped in improving the *accuracy* of the model. Lastly, for the Random Forest (RF) model, both SMOTE and SMOTETomek resampling techniques helped in improving the *accuracy* of the model.

As shown in Table 25, for the high-pressure dataset, sequential SMOTE, SMOTE, and SMOTETomek resampling techniques helped in improving the *accuracy* of the DAGs model. However, for the KNN model, only SMOTE helped in improving the *accuracy* of the model. For both the DT and RF models, SMOTE, SMOTETomek, and SMOTEENN resampling techniques helped in improving the *accuracy* of the model.

Table 24—The overall *accuracy* of the trained models for the low-pressure (592 kPa) dataset.

Datasets	Overall *Accuracy*			
	KNN	DAGs	DT	RF
Sequential SMOTE	N/A	0.86	N/A	N/A
SMOTE	0.75	0.82	0.81	0.81
SMOTETomek	0.75	0.82	0.84	0.81
SMOTEENN	0.72	0.82	0.91	0.63
Before resampling	0.75	0.77	0.84	0.72

Table 25—The overall *accuracy* of the trained models for the high-pressure (2,060 kPa) dataset.

Datasets	Overall *Accuracy*			
	KNN	DAGs	DT	RF
Sequential SMOTE	N/A	0.88	N/A	N/A
SMOTE	0.55	0.91	0.86	0.77
SMOTETomek	0.51	0.90	0.89	0.77
SMOTEENN	0.51	0.83	0.84	0.75
Before resampling	0.53	0.86	0.81	0.69

The micro-averaged *precision* of the trained models for the low-pressure (592 kPa) dataset and for the high-pressure (2,060 kPa) dataset are shown in Tables 26 and 27, respectively.

Table 26—The micro-averaged *precision* of the trained models for the low-pressure (592 kPa) dataset.

Datasets	Micro-averaged *Precision*			
	KNN	DAGs	DT	RF
Sequential SMOTE	N/A	0.86	N/A	N/A
SMOTE	0.75	0.82	0.81	0.81
SMOTETomek	0.75	0.82	0.84	0.81
SMOTEENN	0.72	0.82	0.91	0.62
Before resampling	0.75	0.77	0.84	0.72

Table 27—The micro-averaged *precision* of the trained models for the high-pressure (2,060 kPa) dataset.

Datasets	Micro-averaged *Precision*			
	KNN	DAGs	DT	RF
Sequential SMOTE	N/A	0.88	N/A	N/A
SMOTE	0.55	0.91	0.86	0.77
SMOTETomek	0.51	0.90	0.89	0.77
SMOTEENN	0.51	0.83	0.84	0.75
Before resampling	0.53	0.86	0.81	0.68

The micro-averaged *recall* of the trained models for the low pressure (592 kPa) dataset and for the high pressure (2,060 kPa) dataset are shown in Tables 28 and 29, respectively.

Table 28—The micro-averaged *recall* of the trained models for the low-pressure (592 kPa) dataset.

Datasets	Micro-averaged *Recall*			
	KNN	DAGs	DT	RF
Sequential SMOTE	N/A	0.86	N/A	N/A
SMOTE	0.75	0.82	0.81	0.81
SMOTETomek	0.75	0.82	0.84	0.81
SMOTEENN	0.72	0.82	0.91	0.62
Before resampling	0.75	0.77	0.84	0.72

Table 29—The micro-averaged *recall* of the trained models for the high-pressure (2,060 kPa) dataset.

Datasets	Micro-averaged *Recall*			
	KNN	DAGs	DT	RF
Sequential SMOTE	N/A	0.88	N/A	N/A
SMOTE	0.55	0.91	0.86	0.77
SMOTETomek	0.51	0.90	0.89	0.77
SMOTEENN	0.51	0.83	0.84	0.75
Before resampling	0.53	0.86	0.81	0.68

The micro-averaged F_1-*score* of the trained models for the low-pressure (592 kPa) dataset and for the high-pressure (2,060 kPa) dataset are shown in Tables 30 and 31, respectively.

Table 30—The micro-averaged F_1-*score* of the trained models for the low-pressure (592 kPa) dataset.

Datasets	Micro-averaged F_1-*score*			
	KNN	DAGs	DT	RF
Sequential SMOTE	N/A	0.86	N/A	N/A
SMOTE	0.75	0.82	0.81	0.81
SMOTETomek	0.75	0.82	0.84	0.81
SMOTEENN	0.72	0.82	0.91	0.62
Before resampling	0.75	0.77	0.84	0.72

Table 31—The micro-averaged F_1-*score* of the trained models for the high-pressure (2,060 kPa) dataset.

Datasets	Micro-averaged F_1-*score*			
	KNN	DAGs	DT	RF
Sequential SMOTE	N/A	0.88	N/A	N/A
SMOTE	0.55	0.91	0.86	0.77
SMOTETomek	0.51	0.90	0.89	0.77
SMOTEENN	0.51	0.83	0.84	0.75
Before resampling	0.53	0.86	0.81	0.68

The PFIs of the input features to the trained DAGs models are shown in **Tables 32-35**, where a higher number indicates a higher importance.

As shown in the tables, for the trained DAGs models with the sequential SMOTE, SMOTE, SMOTETomek, or SMOTEENN preprocessing over- and under-sampling techniques, v_{sl} was determined to be the most important feature for classifying the flow patterns for the low-pressure dataset. However, for the high-pressure dataset, H_L was determined to be the most important feature for classifying the flow patterns.

Table 32—The permutation feature importance of the input features to the directed acyclic graphs model of the sequential SMOTE generated datasets.

Datasets	v_{sl} (m/s)	v_{sg} (m/s)	*Pressure* (kgf/cm^2)	*T* (°C)	H_L (%)	Angle (0°, 1°, and 3°)
Low–pressure 592 kPa	0.36	0.05	0	0	0.14	0
High–pressure 2,060 kPa	0.22	0.26	0.12	-0.02	0.33	0.02

Table 33—The permutation feature importance of the input features to the directed acyclic graphs model of the SMOTE generated datasets.

Datasets	v_{sl} (m/s)	v_{sg} (m/s)	*Pressure* (kgf/cm^2)	*T* (°C)	H_L (%)	Angle (0°, 1°, and 3°)
Low–pressure 592 kPa	0.41	0.05	-0.05	-0.05	0.05	0
High–pressure 2,060 kPa	0.28	0.26	0.02	0	0.35	0.07

Table 34—The permutation feature importance of the input features to the directed acyclic graphs model of the SMOTETomek generated datasets.

Datasets	v_{sl} (m/s)	v_{sg} (m/s)	*Pressure* (kgf/cm^2)	*T* (°C)	H_L (%)	Angle (0°, 1°, and 3°)
Low–pressure 592 kPa	0.41	0.05	0	-0.05	0	0
High–pressure 2,060 kPa	0.26	0.24	0.02	0	0.38	0.05

Table 35—The permutation feature importance of the input features to the directed acyclic graphs model of the SMOTEENN generated datasets.

Datasets	v_{sl} (m/s)	v_{sg} (m/s)	*Pressure* (kgf/cm^2)	*T* (°C)	H_L (%)	Angle (0°, 1°, and 3°)
Low–pressure 592 kPa	0.55	0.09	0	0.05	0.05	0
High–pressure 2,060 kPa	0.21	0.22	-0.02	-0.02	0.36	0.05

The number of the examples in each class (flow pattern) for the low-pressure (592 kPa) dataset and for the high-pressure (2,060 kPa) dataset, in addition to the number of the examples in the training datasets before and after applying the over- and under-sampling techniques, are shown in **Tables 36** and **37**, respectively.

Table 36—The number of the examples in each class (flow pattern) for the low-pressure (592 kPa) dataset and the number of the examples in the training datasets before and after applying the over- and under-sampling techniques.

Class (Flow pattern)	No. of examples in the dataset	No. of examples in the training dataset	SMOTE	SMOTETomek	SMOTEENN	Sequential SMOTE
SL	55	42	42	42	21	27
F	18	11	42	42	42	22
DB	15	10	42	42	37	7
SR	11	6	42	42	40	48
A	6	4	42	42	42	36

Table 37—The number of the examples in each class (flow pattern) for the high-pressure (2,060 kPa) dataset and the number of the examples in the training datasets before and after applying the over- and under-sampling techniques.

Class (Flow pattern)	No. of examples in the dataset	No. of examples in the training dataset	SMOTE	SMOTETomek	SMOTEENN	Sequential SMOTE
SL	98	71	71	70	42	61
EB	46	30	71	70	63	30
DB	30	24	71	71	60	24
In-DB	27	22	71	69	60	17
SW-SR	19	8	71	71	55	8
F	12	8	71	71	67	63
SR-A	6	4	71	71	66	20
SS	5	3	71	71	69	36

4.8 Conclusion

In the present study, machine learning models were trained and tested to predict the gas–liquid two-phase flow patterns for pipelines of a large diameter under low- and high-pressure conditions at three different angles (0°, 1°, and 3°). The datasets used in this study were imbalanced; therefore, over- and under-sampling techniques were applied to the datasets before training and testing the predictive models.

For the low-pressure dataset, the best model for classifying the gas–liquid two-phase flow patterns was the decision tree model with the SMOTEENN preprocessing over and under-sampling technique. By contrast, for the high-pressure dataset, the best model for classifying the gas–liquid two-phase flow patterns was the DAGs model with the SMOTE preprocessing over and under-sampling technique.

For the feature importance of the input parameters to the trained DAGs models, v_{sl} was determined as the most important feature in classifying the flow patterns in the low-pressure dataset. For the high-pressure dataset, H_L was determined to be the most important feature in classifying the flow patterns.

To the best of our knowledge, this is the first study to consider evaluating the effect of over- and under-sampling techniques for KNN and tree-based models on imbalanced datasets for gas–liquid two-phase flow flow-patterns classification in pipelines, in addition to the determination of the PFI of the input parameters to the DAGs models.

The built models and their tree representations can be used to further investigate and trace the results; that is, the classifications made by the models. Further, the built models can be extended to classifying the flow patterns under different conditions by including different sets of

features as a feature selection process for further estimations in upstream exploration and production computations. Furthermore, the trained models can easily be integrated into upstream oil and gas simulators by deploying them as web services.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

The built BDTR model is accurate, and it outperforms the most commonly used empirical correlations and the previous machine learning models in predicting P_b and B_{ob} in terms of the mean absolute percentage error. This model brings significant predictive capability and versatility to datasets with multiple missing input features. In addition, it is the most accurate model to date in predicting the bubble point pressure and the oil formation volume factor at the bubble point pressure of crude oils. For the built models, the most important feature in estimating P_b and B_{ob} is R_s.

For Kuwaiti crude oils, our results show that the developed predictive model with the presence of missing data outperforms the most commonly used empirical correlations in predicting P_b in terms of the mean absolute percentage error. In addition, for the built model, the most important feature in estimating P_b is R_s. Furthermore, the dataset used in this study is novel and represents oil wells of the Greater Burgan oil fields in the State of Kuwait.

For liquid holdup, our results show that the proposed BDTR model outperforms the most commonly used empirical correlations and the fuzzy logic model used in estimating H_L in gas–liquid two-phase flow. For the built model, the most important input feature in estimating H_L is the superficial gas velocity (v_{sg}). The empirical correlations developed in the past for identifying H_L in the multiphase flow phenomenon can only be applied under certain flow conditions for

which they were originally developed, but this machine learning model does not suffer from this limitation.

For the gas–liquid two-phase flow-pattern classification, our results show that the classified two-phase gas–liquid flow patterns are in good agreement with the experimental ones. Moreover, superficial liquid velocity (v_{sl}) and H_L were determined as the most important features in classifying the flow patterns.

The interpretability and the heuristic feature importance of the tree-based models can enhance understandability in investigating producing oil and gas wells (not a black-box).

Summary

Tree-based models showed superior performance for estimating:

- BDTR: P_b & B_{ob} (complete and incomplete).
- BDTR: Liquid holdup (complete and incomplete).
- Multiclass decision jungle (DAGs), decision tree and random forest: flow patterns (imbalanced).

Incomplete and imbalanced datasets were considered in this study as real-world scenarios:

- Imputation: for the incomplete datasets.
- Over- and under-sampling: for the imbalanced datasets.

Feature importance:

- P_b and B_{ob}: solution gas–oil ratio (R_s).
- Liquid holdup: superficial gas velocity (v_{sg}).
- Flow patterns: low-pressure dataset: (v_{sl}); high-pressure dataset: (H_L).

5.2 Future work

As an interpretable approach, the evaluated models can be used as an alternative modeling scheme in PVT characterization, gas-liquid two-phase flow liquid holdup and flow patterns studies, where the importance of the input features can be heuristically and accurately determined. This can be applied to preventive maintenance and anomaly detection studies where prediction decisions can be further investigated by the interpretable representation of the decision trees. In addition, based on the results and the interpretability of the models, the physical meaning can further be explored as future work. Moreover, the built models can be retrained by including different sets of features as a feature selection process for further estimations in upstream exploration and production processes, such as in gas–liquid multiphase flow liquid loading studies. Furthermore, the studied models can be integrated into simulators.

The gas-liquid two-phase flow pattern classification will be reconsidered as a state estimation problem of a dynamically changing system (hidden Markov model) rather than a static classification problem, as in the current study. In addition, additional fluid mechanical parameters, factors and conditions that might affect the gas-liquid two-phase flow pattern will be considered.

Acknowledgments

Words cannot express my gratitude to my professors for their invaluable patience, advice, and support. I'm thankful to the defense committee for their generosity in providing me knowledge and expertise.

This research would not have been possible without the support from my friend Ghazi Alosaimi, a team leader at Kuwait Oil Company, who helped me in obtaining the datasets from KOC.

I acknowledge the support of the Kuwait Oil Company Research and Development Department. The authors are especially indebted to Prof. Ridha Gharbi, a senior consultant at KOC, and to Mr. Jassim Al-Kanderi, the Greater Burgan Studies-I team leader, for providing the worldwide crude oil dataset and the Kuwaiti crude oil system dataset. I also acknowledge the support received from the Japan Oil, Gas and Metals National Corporation (JOGMEC).

Lastly, I am grateful for the support received from my family, especially my parents, spouse, brother, and children. Their belief in me has kept my spirits and motivations high during this journey.

Publications

Journals:

- H. Al-Hamadi, I. -R. Chen, D. -C. Wang and M. Almashan, "Attack and Defense Strategies for Intrusion Detection in Autonomous Distributed IoT Systems," IEEE Access, Vol. 8 (2020), pp.168994—169009. (Published)

- Meshal Almashan, Arsalan Zolfaghari, Yoshiaki Narusue, Hiroyuki Morikawa,"Estimating Pressure–Volume–Temperature Properties of Crude Oil Systems Using Boosted Decision Tree Regression," Journal of the Japan Petroleum Institute. (Accepted)

- Meshal Almashan, Yoshiaki Narusue, Hiroyuki Morikawa,"Gas-Liquid Two-Phase Flow Pattern Classification for Imbalanced Datasets Using Machine Learning Techniques," IEEE Access. (To be submitted)

International conferences:

- A. Almutairi and M. Almashan, "Instance Segmentation of Newspaper Elements Using Mask R-CNN," 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, FL, USA, 2019, pp. 1371-1375.

- Meshal Almashan, Yoshiaki Narusue, and Hiroyuki Morikawa, "Estimating PVT Properties of Crude Oil Systems Based on a Boosted Decision Tree Regression Modelling Scheme with K-Means Clustering," SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition. Bali, Indonesia, 2019.

- Meshal Almashan, Yoshiaki Narusue, and Hiroyuki Morikawa, "A Decision Tree Regression Modeling Scheme for Estimating the PVT Properties of Kuwaiti Crude Oil Systems Using Incomplete Datasets," Abu Dhabi International Petroleum Exhibition & Conference. Abu Dhabi, UAE, 2019.

- Meshal Almashan, Yoshiaki Narusue, and Hiroyuki Morikawa, "Decision Tree Regressions for Estimating Liquid Holdup in Two-Phase Gas-Liquid Flows," Abu Dhabi International Petroleum Exhibition & Conference. Abu Dhabi, UAE, 2020.

研究会:

- Meshal Almashan, Yoshiaki Narusue, Hiroyuki Morikawa, “A Study on a Feature Based Clustering and Decision Tree Regressions for Estimating the Bubble Point Pressure of Crude Oils,” IEICE Technical Report, vol. 118, no. 282, ASN2018-68, pp. 75-80, Tokyo Denki University, Tokyo, Japan, Nov 5-6, 2018.

- Meshal Almashan, Yoshiaki Narusue, Hiroyuki Morikawa, “An Experimental Study on Gas-Liquid Two-Phase Flow-Pattern Classification in Gas Wells Using Machine Learning Techniques,” IEICE Technical Report, 信学技報, vol. 120, no. 382, SeMI2020-60, pp. 13-18, Mar. 1-2, 2021, Japan.

全国大会:

- Meshal Almashan, Ramin Khatami, Theerat Sakdejayont, Yoshiaki Narusue, Hiroyuki Morikawa, “Decision Forest Regression for PVT Characterization in the Oil & Gas Upstream Operations,” IEICE society conference, BS-7-14 , September 2018.

- Ramin Khatami, Meshal Almashan (The Univ. of Tokyo) ,Mohsen Jafari Songhori (Univ. of Twente) ,Theerat Sakdejayont, Yoshiaki Narusue, Hiroyuki Morikawa(The Univ. of Tokyo),“Product Ideas Success Prediction in Crowdfunding Platforms Using Textual Similarity Analysis,” IEICE society conference, BS-7-11 , September 2018.

- Meshal Almashan, Yoshiaki Narusue, Hiroyuki Morikawa, “A Universal PVT Characterization: Boosted Decision Tree Regressions for Worldwide Crude Oils Datasets,” IEICE society conference, D-20-20, March 2019.

- Meshal Almashan, Yoshiaki Narusue, Hiroyuki Morikawa, “A Study on Liquid Holdup Prediction in Oil and Gas Wells Using a Decision Tree Regression Predictive Model,” IEICE society conference, B-15-6, September 2019.

Nomenclature

ANN	Artificial neural network
B_o	Oil formation volume factor, rb/stb
B_{ob}	Oil formation volume factor at bubble point pressure, bbl/stb
E_a	Mean absolute percentage error (MAPE), %
E_i%	Relative deviation
EOS	Equation of state
H_L	Slip liquid holdup, %
n	Total number of data points
P_b	Bubble point pressure, psi
PFI	Permutation feature importance, score
PVT	Pressure–volume–temperature
R_s	Solution gas–oil ratio, scf/stb
STB	Stock-tank barrel
T	Temperature, °F/°C
v_{sg}	Superficial gas velocity, ft/sec (m/sec)
v_{sl}	Superficial liquid velocity, ft/sec (m/sec)
γ_{API}	Oil API gravity, ^{o}API
γ_g	Gas specific gravity, air = 1.0
γ_o	Oil-specific gravity, air = 1.0
θ	Inclination angle, degree

References

Abduvayt, P. (2003) *Experimental and modeling studies for gas-liquid two-phase flow in slightly inclined pipes at low- and high-pressure conditions*, Ph.D. Dissertation, Waseda University, Japan.

Abduvayt, P. *et al.* (2003a) "Experimental and modeling studies for gas-liquid two-phase flow at high pressure conditions," *Journal of the Japan Petroleum Institute*, 46(2), pp. 111–125, [online] doi:10.1627/jpi.46.111.

Abduvayt, P., Manabe, R. and Arihara, N. (2003b) "Effects of pressure and pipe diameter on gas-liquid twophase flow behavior in pipelines," *SPE Annual Technical Conference and Exhibition*, pp. SPE-84229-MS, [online] doi:10.2523/84229-MS.

Ahmed, T. (1989) *Hydrocarbon phase behavior*. Houston, Texas: Gulf Publishing.

Al-Marhoun, M. A. (1988) "PVT correlations for Middle East crude oils," *Journal of Petroleum Technology*, 40(5), p. 650.

Al-Marhoun, M. A. (1992) "New correlations for formation volume factors of oil and gas mixtures," *Journal of Canadian Petroleum Technology*, 31(3), p. 22.

Almashan, M. *et al.* (2019a) "Estimating PVT properties of crude oil systems based on a Boosted Decision Tree Regression modelling scheme with K-Means Clustering," *The SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition*.

Almashan, M. *et al.* (2019b) "A decision tree regression modeling scheme for estimating the PVT properties of Kuwaiti crude oil systems using incomplete datasets," *The Abu Dhabi International Petroleum Exhibition & Conference, Society of Petroleum Engineers*.

Almashan, M. *et al.* (2020) "Decision tree regressions for estimating liquid holdup in two-phase gas-liquid flows," *Abu Dhabi International Petroleum Exhibition & Conference*.

Almehaideb, R. A. (1997) "Improved PVT correlations for UAE crude oils," *Middle East Oil Show and Conference*.

Al-Safran, E. and Al-Qenae, K. (2018) "A study of flow-pattern transitions in high-viscosity oil-and-gas two-phase flow in horizontal pipes," *SPE Production & Operation*, 33(2), [online] doi:10.2118/187939-PA.

Al-Shammasi, A. A. (2001) "A review of bubblepoint pressure and oil formation volume factor correlations", *SPE Reservoir Evaluation & Engineering*, 4(2), p. 146.

Arihara, N. (2002) *Study on improvement of production efficiency. Report of entrustment study on prediction model of gas-liquid two-phase flow pattern in inclined pipes*, Dissertation, Waseda University, Japan.

Arya, A. *et al.* (1981) "Comparison of two phase liquid holdup and pressure drop correlations across flow regime boundaries for horizontal and inclined pipes," *SPE Annual Technical Conference and Exhibition*.

Baker, O. (1954) "Simultaneous flow in oil and gas," *Oil and Gas Journal*, 53, pp. 185–195.

Barnea, D. *et al.* (1982) "Flow pattern transition for downward inclined two phase flow;

horizontal to vertical," *Chemical Engineering Science*, 37(5), pp. 735–740.

Barnea, D. (1987) "A unified model for predicting flow-pattern transitions for the whole range of pipe inclinations," *International journal of multiphase flow*, 13(1), p. 1.

Beggs, D. H. and Brill, J. P. (1973) "A study of two-phase flow in inclined pipes," *Journal of Petroleum Technology*, 25(5), pp. 607–617, [online] doi:10.2118/4007-PA.

Bhagwat, S. M. *et al.* (2016) "Experimental investigation of non-boiling gas-liquid two phase flow in upward inclined pipes," *Experimental Thermal and Fluid Science*, 79, p. 301.

Brill, J. P. and Mukherjee, H. (1999) *Multiphase flow in wells*. Texas: Society of Petroleum Engineers Richardson.

Brito, A. *et al.* (2015) "Flow pattern transition model for gas highly viscous fluids in horizontal pipelines," *17th International Conference on Multiphase Production Technology*.

Crawford, T. J. *et al*. (1985) "Two-phase flow patterns and void fractions in downward flow Part I: Steady-state flow patterns," *International journal of multiphase flow*, 11(6), p. 761.

Dabirian, R. *et al.* (2018) "The effects of phase velocities and fluid properties on liquid holdup under gas-liquid stratified flow," *SPE Western Regional Meeting.*

De Ghetto, G., Paone, F. and Villa, M. (1995) "Pressure-volume-temperature correlations for heavy and extra heavy oils," *SPE International Heavy Oil Symposium*, [online] doi:10.2118/30316-MS.

He, D. *et al.* (2021). "Gas–Liquid Two-Phase Flow Pattern Identification of a Centrifugal Pump Based on SMOTE and Artificial Neural Network," *Micromachines*, 13(1), p. 2.

Dindoruk, B. *et al.* (2004) "PVT properties and viscosity correlations for Gulf of Mexico oils.," *SPE Reservoir Evaluation & Engineering*, 7(6), p. 427.

Dokla, M. E. *et al.* (1992) "Correlation of PVT properties for UAE crudes," *SPE Formation Evaluation*, 7(1), p. 41.

El-Sebakhy, E. (2009) "Forecasting PVT Properties of Crude Oil Systems Based on Support Vector Machines Modeling Scheme," *Journal of Petroleum Science and Engineering*, 64(1–4), p. 25.

El-Sebakhy, E. A. *et al.* (2007) "Support vector machines framework for predicting the PVT properties of crude oil systems," *SPE Middle East Oil and Gas Show and Conference*.

Elsharkawy, A. M. (1998) "Modeling the properties of crude oil and gas systems using RBF network," *SPE Asia Pacific Oil and Gas Conference and Exhibition*.

Elsharkawy, A. M. *et al.* (1995) "Assessment of the PVT correlations for predicting the properties of Kuwaiti crude oils," *Journal of Petroleum Science and Engineering*, 13(3-4), p. 219.

Farshad, F. F. *et al.* (1992) "Empirical PVT Correlations for Colombian Crudes," *Unsolicited SPE Paper*, p. 24538.

Fath, H. A. *et al.* (2018) "Development of an Artificial Neural Network Model for Prediction of Bubble Point Pressure of Crude Oils," *Petroleum*, 4(3), p. 281.

Frashad, F. *et al.* (1996) “Empirical PVT correlations for Colombian crude oils,” *SPE Latin America/Caribbean Petroleum Engineering Conference*.

Freecodecamp.org (2020) *Random forest classifier tutorial: How to use tree-based algorithms for machine learning*. Available at: https://www.freecodecamp.org/news/how-to-use-the-tree-based-algorithm-for-machine-learning/ (Accessed: 12 January 2021).

Gharbi, R. B. *et al.* (1997) “Neural Network Model for Estimating the PVT Properties of Middle East Crude Oils,” *Middle East Oil Show and Conference Society of Petroleum Engineers*.

Gharbi, R. B. *et al.* (1999) “Universal Neural-Network-Based Model for Estimating the PVT Properties of Crude Oil Systems,” *Energy Fuel*, 13(2), p. 454.

Gharbi, R. B. *et al.* (2003) “Predicting the Bubble-Point Pressure and Formation-Volume-Factor of Worldwide Crude Oil Systems.,” *Journal of Petroleum Science and Technology*, 21(1–2), p. 53.

Glasø, O. (1980) “Generalized pressure–volume–temperature correlations,” *Journal of Petroleum Technol*ogy, 32(5), pp. 785–795.

Goda, M. H. *et al.* (2003) “Prediction of the PVT Data Using Neural Network Computing Theory,” *27th Annual SPE International Technical Conference and Exhibition in Abuja*.

Gokcal, B. *et al.* (2010) “Prediction of slug frequency for high-viscosity oils in horizontal pipes,” *SPE Projects, Facilities & Construction*, 5(3), pp. 136–14, [online] doi:10.2118/124057-MS.

Gould, T. L., Tek, M. R. and Katz, D. L. (1974) “Two-phase flow through vertical, inclined, or curved pipe,” *Journal of Petroleum Technology*, 26(8), pp. 915–926, [online] doi:10.2118/4487-PA.

Guzhov, A. I., Mamayev, V. A. and Odishariya, G. E (1967) “A study of transportation in gas-liquid systems,” *Proceedings of the 10th Insights into Gas Conference*.

Hadavimoghaddam, F. *et al.* (2019) “Oil and gas properties (PVT) correlation using neural network,” *Oil & Gas Journal*, 07, p. 104.

Hajizadeh, Y. (2007) “Production and Operations Symposium,” *Proceedings of the Society of Petroleum Engineers.*

Hanyang, G. U. *et al.* (2007) “Stability of stratified gas-liquid flow in horizontal and near horizontal pipes,” *Chinese Journal of Chemical Engineering*, 15(5), p. 619.

He, D. *et al.* (2021) “Gas–liquid two-phase flow pattern identification of a centrifugal pump based on SMOTE and artificial neural network,” *Micromachines*, 13(1), p. 2.

Hughmark, G.A. *et al.* (1961) “Holdup and pressure drop with gas-liquid flow in a vertical pipe,” *AIChE Journal*, 7(4), p. 677.

Imbalanced-learn.org (2021) *Combination of over- and under-sampling*. Available at: https://imbalanced-learn.org/stable/combine.html (Accessed: 12 January 2021).

Jacobson, H. A. (1967) “The effect of nitrogen on reservoir fluid saturation pressure,” *Journal of Canadian Petroleum Technology*, 6(3), p. 101.

James, P. *et al.* (1999) “Multiphase flow in wells,” *Proceedings of the Society of Petroleum*

Engineers.

Jarrahian, A. *et al.* (2015) "Empirical Estimating of Black Oils Bubblepoint (Saturation) Pressure," *Journal of Petroleum Science and Engineering*, 126, p. 69.

Kang, C. and Jepson, W. P. (2002) "Flow regime transitions in large diameter inclined multiphase pipelines," *NACE International Corrosin*, p. 02243.

Kartoatmodjo, T. *et al.* (1994) "Large Data Bank Improves Crude Physical Property Correlations," *Oil & Gas Journal*, 92(27), p. 51.

Khairy, M. *et al.* (1998) "PVT Correlations Developed for Egyptian Cr," *Oil & Gas Journal*, 96(19), p. 114.

Khaledi, H. A. *et al.* (2014) "Investigation of two-phase flow pattern, liquid holdup and pressure drop in viscous oil-gas flow," *International Journal of Multiphase Flow*, 67, pp. 37–51.

Khoukhi, A. *et al.* (2011) "Support Vector Regression and Functional Networks for Viscosity and Gas Oil Ratio Curves Estimation," *International Journal of Computational Intelligence Application*, 10(3), p. 269.

Kjolaas, J. *et al.* (2015) "Experiments for low liquid loading with liquid holdup discontinuities in two-and three-phase flows," *17th International Conference on Multiphase Production Technology*.

Labedi, R. M. (1990) "Use of Production Data to Estimate the Saturation Pressure, Solution GOR, and Chemical Composition of Reservoir Fluids," *SPE Latin America Petroleum Engineering Conference*.

Lasater, J. (1958) "Bubble Point Pressure Correlation," *Journal of Petroleum Technology*, 10(5), p. 65.

Li, X. *et al.* (2014) "Multiphase flow pattern recognition in horizontal and upward gas-liquid flow using support vector machine models," *SPE Annual Technical Conference and Exhibition*.

Macary, S. *et al.* (1993) "Derivation of PVT Correlations for the Gulf of Suez Crude Oils," *Sekiyu Gakkai Shi* (J. Jpn. Petrol. Inst.), 36(6), p. 472.

Machinelearningmastery.com (2020a) *How to calculate precision, recall, and f-measure for imbalanced classification.* Available at: https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/ (Accessed 19 January 2021).

Machinelearningmastery.com (2020b) *How to combine over-sampling and undersampling for imbalanced classification*. Available at: https://machinelearningmastery.com/combine-over-sampling-and-undersampling-for-imbalanced-classification/ (Accessed 12 January 2021).

Malallah, A. M., Gharbi, R. and Algharaib, M. (2006) "Accurate estimation of the world crude oil PVT properties using graphical alternating conditional expectation," *Energy Fuels*, 20(2), pp. 688–698, [online] doi:10.1021/ef0501750.

Mandhane, J. M. *et al.* (1974) "A Flow pattern map for gas-liquid flow in horizontal pipes: Predictive models," *International Journal of Multiphase Flow*, 1, pp. 537–553.

Mask, G., Wu, X. and Ling, K. (2019) "An improved model for gas-liquid flow pattern prediction based on machine learning," *Journal of Petroleum Science and Engineering*, 183, p. 106370.

McCain, W. D. J. (1990) *The properties of petroleum fluids*. Tulsa, Oklahoma: PennWell Books.

McCain, W. D. J. *et al.* (1998) "Correlation of Bubblepoint Pressures for Reservoir Oils—a Comparative Study," *SPE Eastern Regional Meeting*.

Microsoft.com (2019a) *Boosted Decision Tree Regression module*. Available at: https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/boosted-decision-tree-regression (Accessed: 21April 2021).

Microsoft.com (2019b) *Decision forest regression.* Available at: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/decision-forest-regression. (Accessed: 22 April 2021).

Microsoft.com (2019c) *Multiclass decision jungle*. Available at: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-decision-jungle. (Accessed: 30 April 2021).

Microsoft.com (2019d) *Permutation feature importance*. Available at: https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/permutation-feature-importance (Accessed: 22 April 2021).

Minami, K. and Brill, J. P. (1987) "Liquid holdup in wet-gas pipelines," *SPE PE*, 36, [online] doi:10.2118/14535-PA.

Moghadassi, A.R. *et al.* (2009) "A New Approach for Estimation of PVT Properties of Pure Gases Based on Artificial Neural Network Model," *Brazil Journal of Chemical Engineering*, 26, p. 199.

Mukherjee, C. *et al.* (1983) "Liquid holdup correlations for inclined two-phase flow," *Journal of Petroleum Technology*, 35(5), pp. 1–3.

Naseri, A. *et al.* (2005) "A Correlation Approach for Prediction of Crude Oil Viscosities," *Journal of Petroleum Science and Engineering*, 47(3–4), p. 163.

Obomanu, D. A. *et al.* (1987) "Correlating the PVT properties of Nigerian crudes," *Journal of Energy Resources Technology*, 109, p. 214.

Oddie, G. *et al.* (2003) "Experimental study of two and three phase flows in large diameter inclined pipes," *International Journal of Multiphase Flow*, 29(4), pp. 527–558, [online] doi:10.1016/S0301-9322(03)00015-6.

Olatunji, S. O. *et al.* (2015) "Modeling permeability and PVT properties of oil and gas reservoir using hybrid model based on type-2 fuzzy logic systems," *Neurocomputing*, 157, p. 125.

Oloso, M. A. *et al.* (2017) "Hybrid Functional Networks for Oil Reservoir PVT Characterization," *Expert Systems with Applications*, 87, p. 363.

Omar, M. I. *et al.* (1993) "Development of New Modified Black Oil Correlations for Malaysian Crudes," *SPE Asia Pacific Oil and Gas Conference*.

Omebere-Iyari, N. K., Azzopardi, B. J. and Ladam, Y. (2007) “Two-phase flow patterns in large diameter vertical pipes at high pressures,” *American Institute of Chemical Engineers Journal*, 53(1), pp. 2493–2504, [online] doi:10.1002/aic.11288.

Osman, E. A. *et al.* (2001) “Prediction of Oil PVT Properties Using Neural Networks,” *SPE Middle East Oil Show Society of Petroleum Engineers*.

Osman, E. *et al.* (2005) “Artificial Neural Networks Models for Predicting PVT Properties of Oil Field Brines,” *14th SPE Middle East Oil & Gas Show and Conference*.

Palmer, C. M. (1975) *Evaluation of inclined pipe two-phase liquid holdup correlations using experimental data*, MS Thesis, University of Tulsa.

Payne, G. A. *et al.* (1979) “Evaluation of inclined-pipe, two-phase liquid holdup and pressure-loss correlation using experimental data (includes associated paper 8782),” *Journal of Petroleum Technology*, 31(9), pp. 1–198, [online] doi:10.2118/6874-PA.

Perry, R. and Green, H. (1999) *Perry’s chemical engineers’ handbook*. Seventh edition. New York, New York: McGraw-Hill.

Petrosky Jr., G. E. *et al.* (1998) “Pressure–Volume–Temperature Correlations for Gulf of Mexico Crude Oils,” *SPE Reservoir Evaluation & Engineering*, 1(5), p. 416.

Petrosky, G. E. and Farshad, F. F. (1993) “Pressure-volume-temperature correlations for Gulf of Mexico Crude Oils,” *SPE Annual Technical Conference and Exhibition*, [online] doi:10.2118/26644-MS.

Rafiee-Taghanaki, S. S. *et al.* (2013) “Implementation of SVM Framework to Estimate PVT Properties of Reservoir Oil,” *Fluid Phase Equilibria*, 346, pp. 25–32.

Ramirez, A. M. *et al.* (2017) “Prediction of PVT properties in crude oil using machine learning techniques MLT,” *SPE Latin America and Caribbean Petroleum Engineering Conference*.

Rammay, M., Hussain, G. and Al-Nuaim, N. (2015) “Flow regime prediction using fuzzy logic and modification in Beggs and Brill multiphase correlation,” *International Petroleum Technology Conference*.

Rodrigues, S. *et al.* (2019) “Pressure effects on low-liquid-loading oil/gas flow in slightly upward inclined pipes: flow pattern, pressure gradient, and liquid holdup,” *SPE Journal*, 24(5), pp. 2–221.

Ross, T. J. (2004) *Fuzzy logic with engineering applications*. 2nd ed. New York, NY: John Wiley & Sons.

Rudin, C. (2019) “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, 1, pp. 206–215.

Sami, N. and Aqqour, M. (1992) “Evaluation of horizontal multiphase flow correlations,” *Middle East Oil Show*.

Shoham, O (1982) *Flow Pattern Transition and Characterization in Gas-Liquid Two-Phase Flow in Inclined Pipes*, Ph.D. Dissertation, Tel-Aviv University, Israel.

Shoham, O. *et al.* (1984) “Stratified turbulent-turbulent gas-liquid flow in horizontal and inclined pipes,” *AIChE journal*, 30(3), p. 377.

Shotton, J. *et al.* (2013) “Decision jungles: Compact and rich models for classification,” *Advances in Neural Information Processing Systems*, 26, pp. 234–242.

Standing, M. B. (1947) “A Pressure–Volume–Temperature Correlation for Mixtures of California Oils and Gases,” *Drilling and Production Practice*, 275.

Standing, M. B. (1977) “Volumetric and phase behavior of oil field hydrocarbon systems,” *Proceedings of the Society of Petroleum Engineers of AIME*.

Standing, M. B. (1962) “Oil-system correlations,” in Frick, T. C. (ed.) Handbook. Texas: SPE Richardson, Chap. 19.

Taitel, Y. and Dukler, A. E (1976) “A model for predicting flow regime transitions in horizontal and near horizontal gasliquid flow,” *AIChE Journal*, 22(1), pp. 47–55.

Talebi, R. *et al.* (2014) “Application of soft Computing Approaches for Modeling Saturation Pressure of Reservoir Oils,” *Journal of Natural Gas Science and Engineering*, 20(8).

Vasquez, M. and Beggs, H. D. (1980) “Correlations for Fluid Property Predictions,” *Journal of Petroleum Technology Endpoint kro Kro*, 8.

Vieira, C. *et al.* (2018) “Experimental investigation of twophase flow regime in an inclined pipe,” *11th North American Conference on Multiphase Production Technology*.

Wen, Y. *et al.* (2017) “Experimental study of liquid holdup of liquid-gas two-phase flow in horizontal and inclined pipes,” *International Journal of Heat and Technology,* 35(4), pp. 713–720, [online] doi:10.18280/ijht.350404.

APPENDIX

A. APPENDIX A

Table A-1—Statistical measurements of the multiregional pressure–volume–temperature datasets (complete).

Property	Min	Max	Mean
T (°F)	74	342	184.2
R_s (scf/stb)	9	3370	496
γ_g (air = 1)	0.5	1.67	0.8
γ_{API} (°API)	14.3	59	36.82
P_b (psi)	79	7130	1644.4
B_{ob} (rb/stb)	1.02	2.92	1.33

Table A-2—Statistical measurements of the regional pressure–volume–temperature datasets (complete).

Property	Min	Max	Mean
T (°F)	112	143	132
R_s (scf/stb)	45	701	413
γ_g (air = 1)	0.84	1.23	1.02
γ_{API} (°API)	9.7	41.5	28.3
P_b (psi)	313.73	1940	1415.2

Table A-3—Hyperparameter settings for the random forest regression model.

	Hyperparameters			
Model	Number of decision trees	Max depth	Min number of samples per leaf node	Number of random splits per node
Random forest regression for P_b (worldwide dataset)	2000	16	1	1024
Random forest regression for B_{ob} (worldwide dataset)	100	16	2	1024

Table A-4—Hyperparameter settings for the boosted decision tree regression model for P_b and B_{ob}.

	Hyperparameters			
Model	Number of decision-trees	Max number of leaves per tree	Min number of samples per leaf node	Learning rate
BDTR for P_b (complete worldwide dataset)	790,000	10	1	0.01
BDTR for B_{ob} (complete worldwide dataset)	9,200	15	1	0.04
BDTR for P_b (incomplete worldwide dataset)	10,000	20	1	0.1
BDTR for B_{ob} (incomplete worldwide dataset)	30,000	12	2	0.01
BDTR for P_b (complete Kuwaiti dataset)	4,000	10	1	0.02
BDTR for P_b (incomplete Kuwaiti dataset)	20	37	2	0.47

Table A-5—The regional dataset representing Kuwaiti crude oil systems.

T (°F)	R_s (scf/stb)	γ_{API} (°API)	γ_g (air = 1)	P_b (psi)
132	469	31.3	1.028	1399
133.5	456	29.45	1.028	1414
134.5	435	28.61	1.014	1450
131	457	31.9	0.993	1395
129	429.7	31.3	1.012	1565
136	607	29.66	0.965	1860
136	621	30	0.985	1839
141	517	29.85	1.006	1715
124	403	29.04	1.086	1230
130.47	701	37.45	1.016	1619
134.64	612	32.93	0.968	1820
136.2	678	34.24	0.972	1860
143	513	30.4	0.963	1739
136	420.7	30.4	1.003	1633
133	537	27.5	0.991	1817
135	463	27.67	0.994	1608
136	489	29.3	1.007	1687
131	491	28.9	1.017	1560
133	554	30.6	0.979	1759
130	527	29.84	1.002	1673
135.7	282	27.7	1.058	950
133.7	387	27.9	1.001	1300
140	496	27.85	1.01	1725

135	504	29.5	1.002	1631
133	432	28.1	1.021	1424
130	484	28.4	1.028	1590
133	495	29	0.989	1708
131.7	437	28.7	0.987	1920
142	191	23.3	1.1159	718.7
132	402	29.47	1.052	1281
132	326	27.5	1.091	1161
132	220	23.65	1.064	869
125	316	26.95	1.147	978
126	205	20.16	1.073	869
126	206	26.42	1.155	650
136	214	28.2	1.147	615
136	164	26.7	1.108	644
134	433	29.11	1.034	1458
134	430	28.93	1.061	1445
137.4	237	22.3	1.09	882
127.7	97	21.1	1.227	419
127.5	74	16.4	1.107	468
132.23	554	31.9	0.999	1585
133	502	30.6	0.976	1641
132	532	32.9	1.0475	1622.73
133	392	24.5	1.043	1544.73
123	260	27.9	0.976	1780
132	457	30.6	0.986	1550
131.5	499	31.5	0.968	1648
131	425	29.5	0.976	1529
130.5	322	33.1	1.057	1085
131	377	30.4	1.02	1199
135	416	30.1	1.017	1534
135	399	25.8	1.005	1550
135	260	26.9	0.975	1225
129.5	325	30.9	1.004	1436
130.6	390	29.7	0.974	1463
132	338	24.4	0.972	1554
127.6	476	29.8	0.997	1746
119.5	598	41.5	1.035	1555
123	574	36.6	1.015	1684
125.3	471	28.6	0.994	1781
137	455	31.5	0.996	1572
133	436	29.8	0.989	1593
135	412	30.5	0.964	1503
132	440	28	0.978	1760
138	454	29.4	1.012	1612
133	422	29.6	1.001	1569
137	434	30.1	1.139	1531
132	426	32.5	1.092	1426
135	489	30.3	0.984	1491
140.4	488	29.1	0.974	1654
140	473	31.3	1.005	1793
141	493	32.1	0.98	1940
138	417	31.9	1.006	1542
125	263	29.1	1.093	815
127	205	22.3	1.004	812
130	608	32.65	1.019	1662

112	595	35.5	0.985	1560
112	570	34.4	0.994	1485
127.5	588	31.9	1.03	1605
131.6	497	29.7	1.044	1605
132.5	194	24.5	1.091	685
133	193	19.9	1.032	900
129	364	32.2	1.009	1100
129	268	26.4	1.043	915
126	673	33.4	0.997	1821
130	549	31.2	1.004	1765
132	509	28.2	1.007	1715
131	499	30.4	1.015	1585
138	93	23.3	1.192	313.73
137	561	32.37	1.026	1728
130	495	34.87	1.094	1442
134	417	29.8	1.042	1352
134	421	30.1	1.035	1365
132	160	20.59	1.033	748
134	444	26.64	0.965	1839
134	464	28.71	0.967	1792
135.9	390	15	0.987	1827
136	48	16.7	0.88	342
136	45	15.47	0.843	346
125	420	30.81	1.043	1494
130	460	30.15	1.033	1511
132	450	30.63	1.01	1570
131.3	240	26	1.037	969
137	364	15.74	0.918	1875
125	498	33.23	1.048	1504
124.4	476	29.79	1.038	1634
137.9	487	32.27	1.054	1441
135	482	31.02	1.026	1428
138	463	9.68	1.176	1767
130	502	29.79	1.048	1623
132	447	25.76	1.015	1688
131	364	17.4	1.0636	1660
135.9	353	14.02	0.995	1769
133	205.2	24.2	1.1041	768
133	133.1	19.34	0.9857	707

Table A-6—Hyperparameter settings for the boosted decision tree regression model for H_L.

	Hyperparameters			
Model	Number of decision-trees	Max number of leaves per tree	Min number of samples per leaf node	Learning rate
BDTR for H_L (complete dataset)	80	20	1	0.1
BDTR for H_L (incomplete dataset)	49270	14	1	0.34

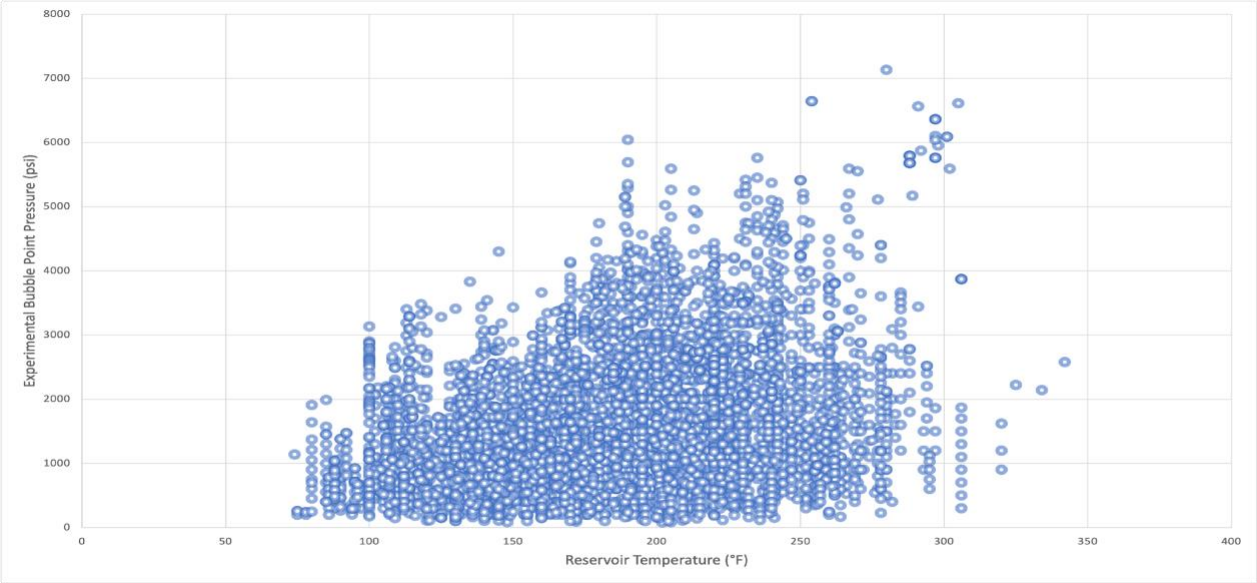

Figure A-1—Cross plot of the reservoir temperature against the experimental bubble point pressure for the worldwide dataset.

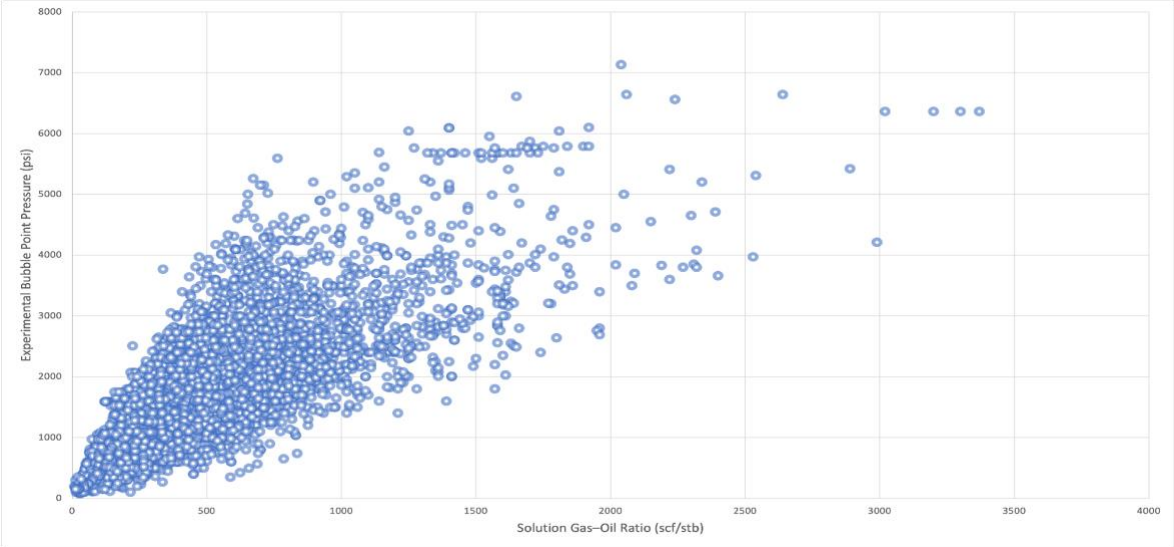

Figure A-2—Cross plot of the solution gas–oil ratio against the experimental bubble point pressure for the worldwide dataset.

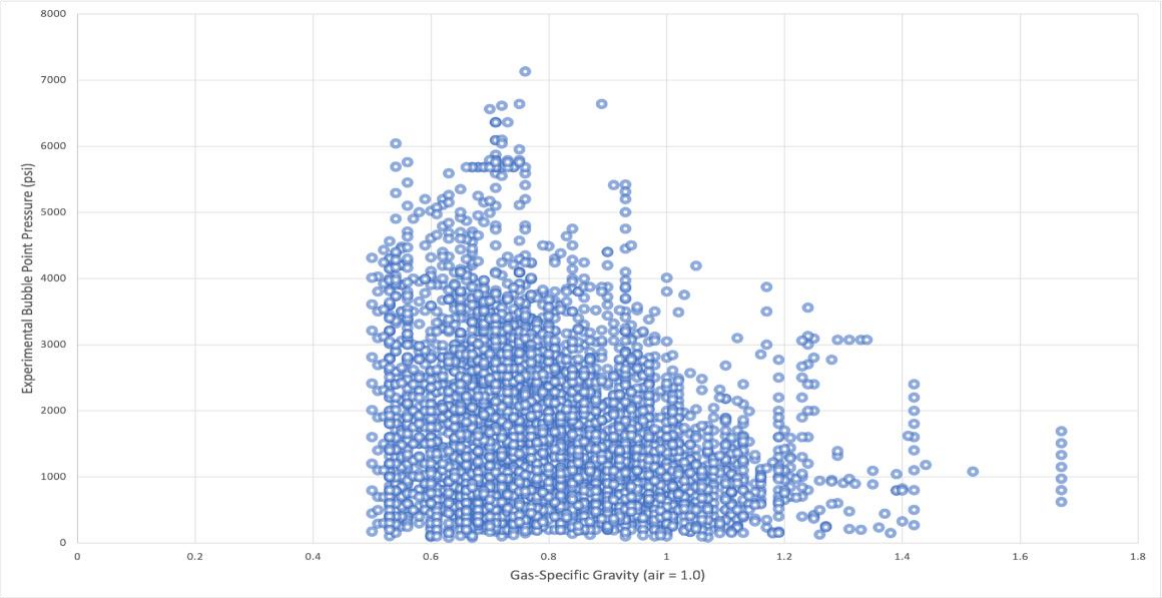

Figure A-3—Cross plot of the gas specific gravity against the experimental bubble point pressure for the worldwide dataset.

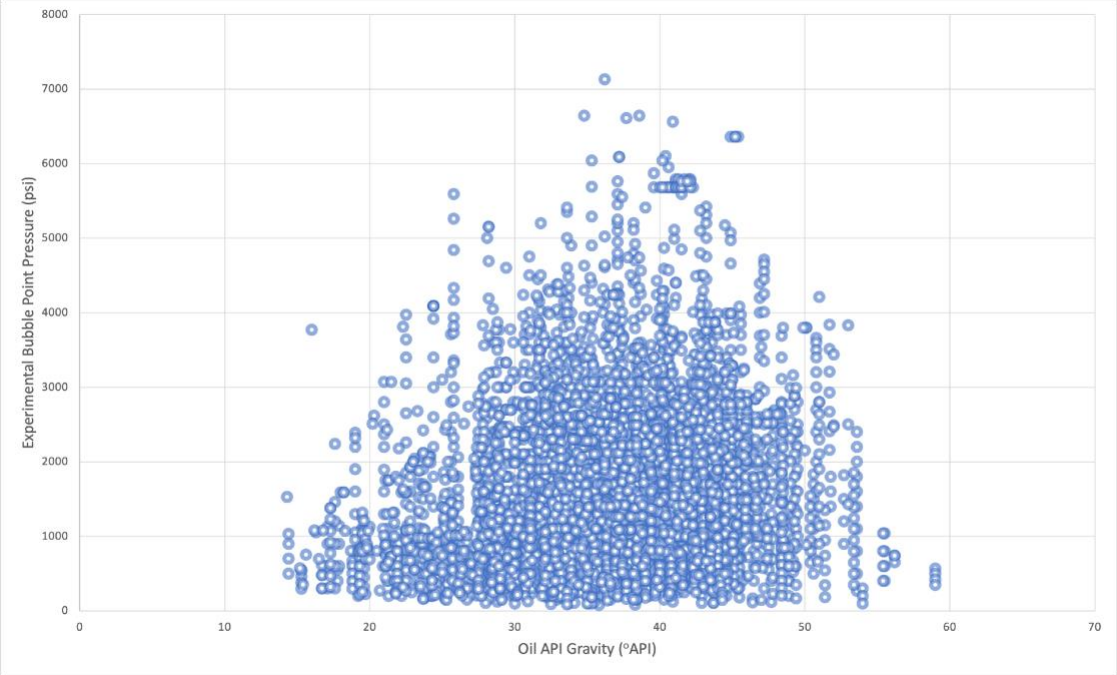

Figure A-4—Cross plot of the oil API gravity against the experimental bubble point pressure for the worldwide dataset.

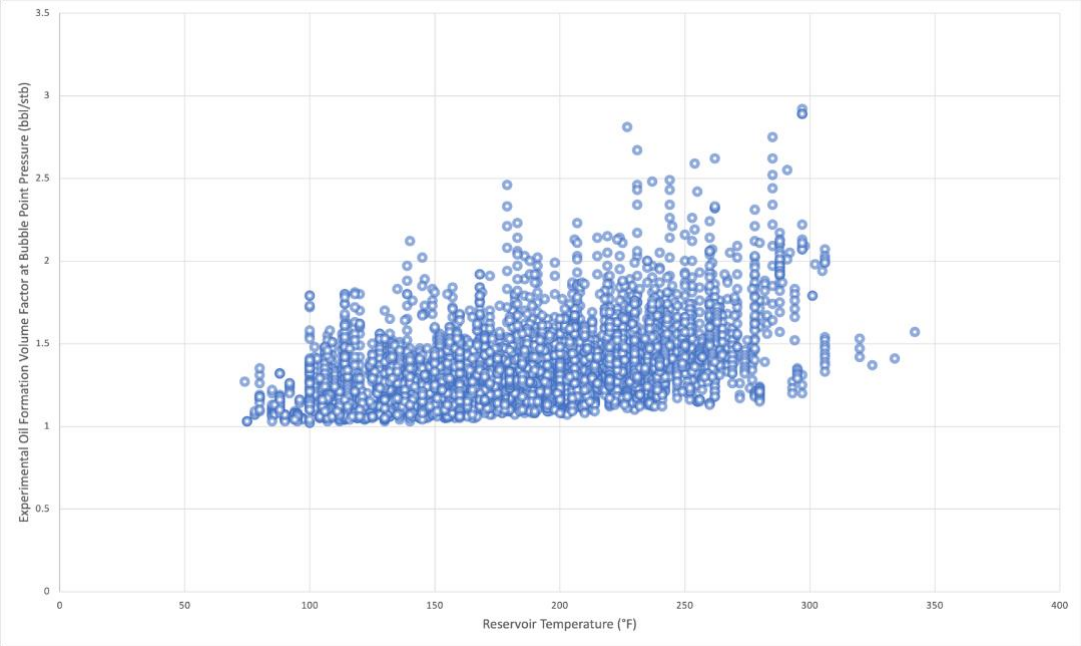

Figure A-5—Cross plot of the reservoir temperature against the experimental oil formation volume factor at bubble point pressure for the worldwide dataset.

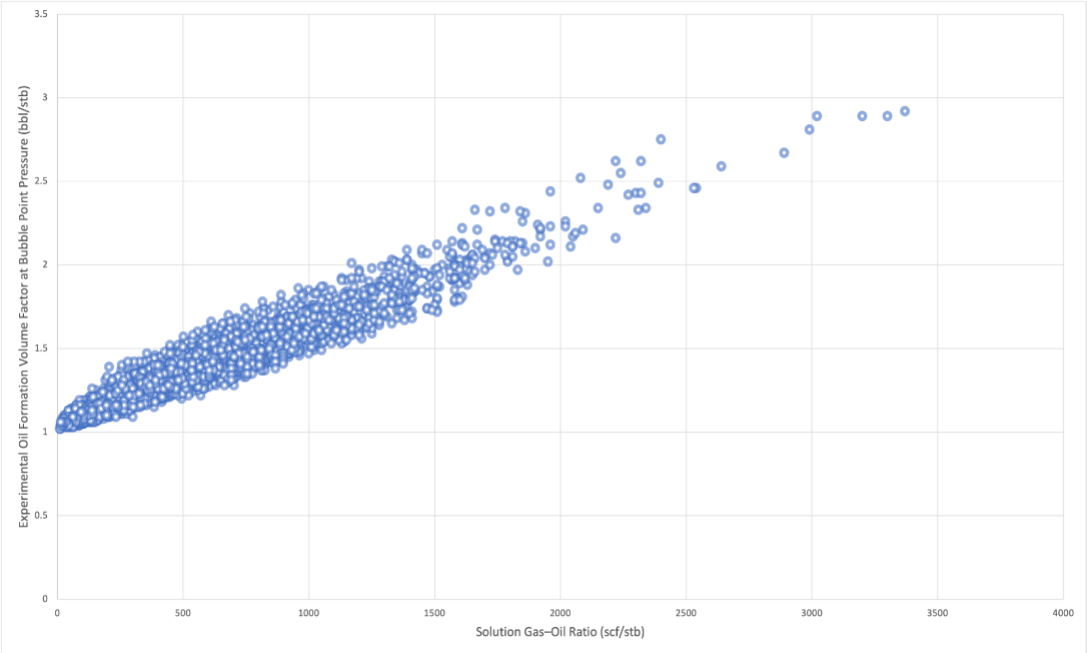

Figure A-6—Cross plot of the solution gas–oil ratio against the experimental oil formation volume factor at bubble point pressure for the worldwide dataset.

Figure A-7—Cross plot of the gas specific gravity against the experimental oil formation volume factor at bubble point pressure for the worldwide dataset.

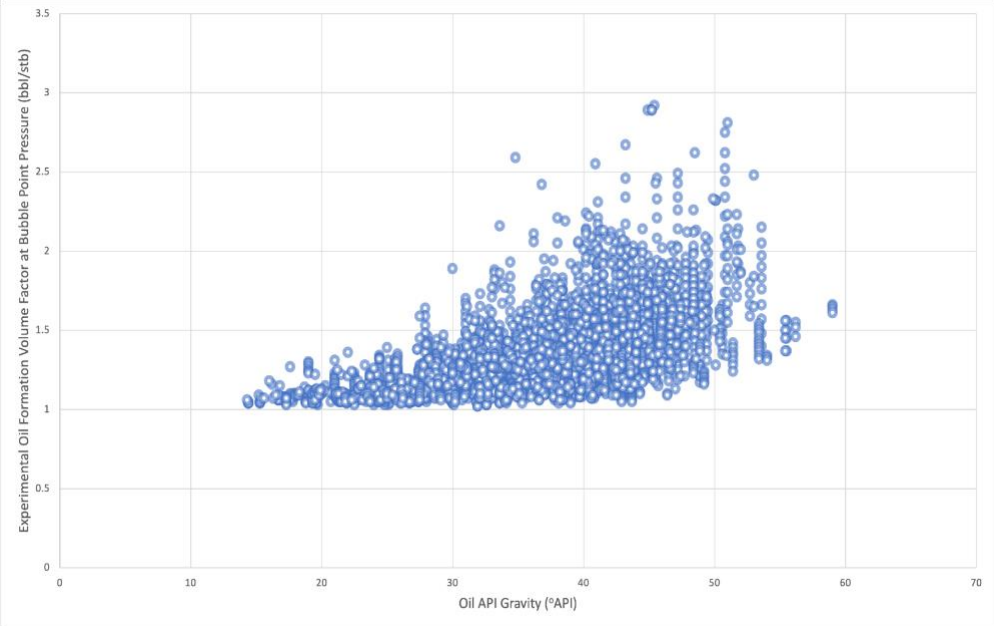

Figure A-8—Cross plot of the oil API gravity against the experimental oil formation volume factor at bubble point pressure for the worldwide dataset.

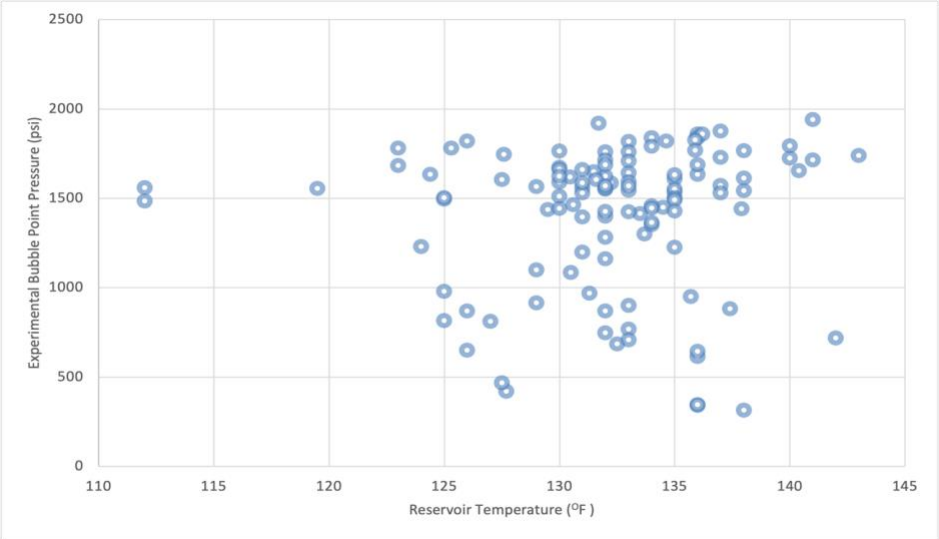

Figure A-9—Cross plot of the reservoir temperature against the experimental bubble point pressure for the Kuwaiti crude oil dataset.

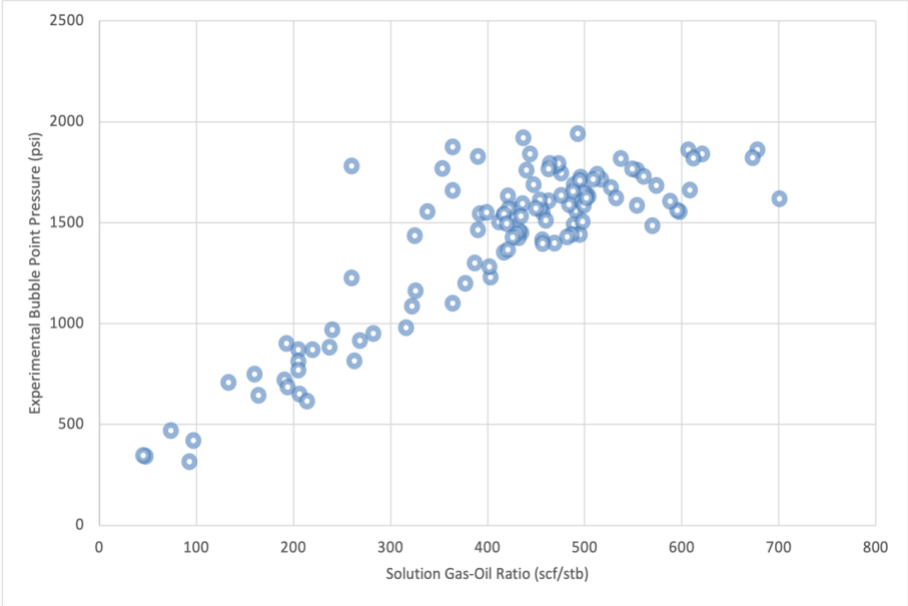

Figure A-10—Cross plot of the solution gas–oil ratio against the experimental bubble point pressure for the Kuwaiti crude oil dataset.

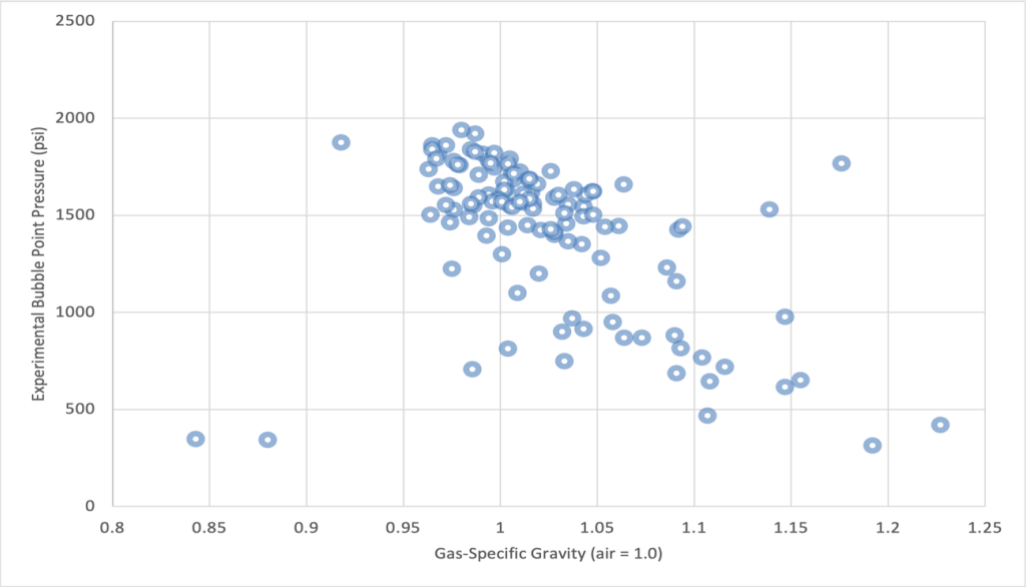

Figure A-11—Cross plot of the gas specific gravity against the experimental bubble point pressure for the Kuwaiti crude oil dataset.

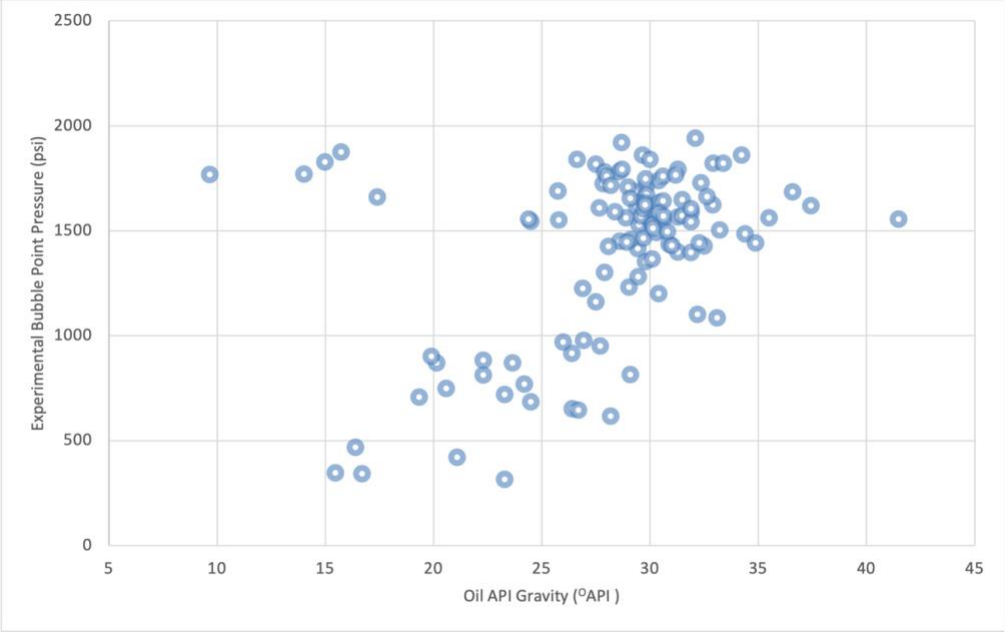

Figure A-12—Cross plot of the oil API gravity against the experimental bubble point pressure for the Kuwaiti crude oil dataset.

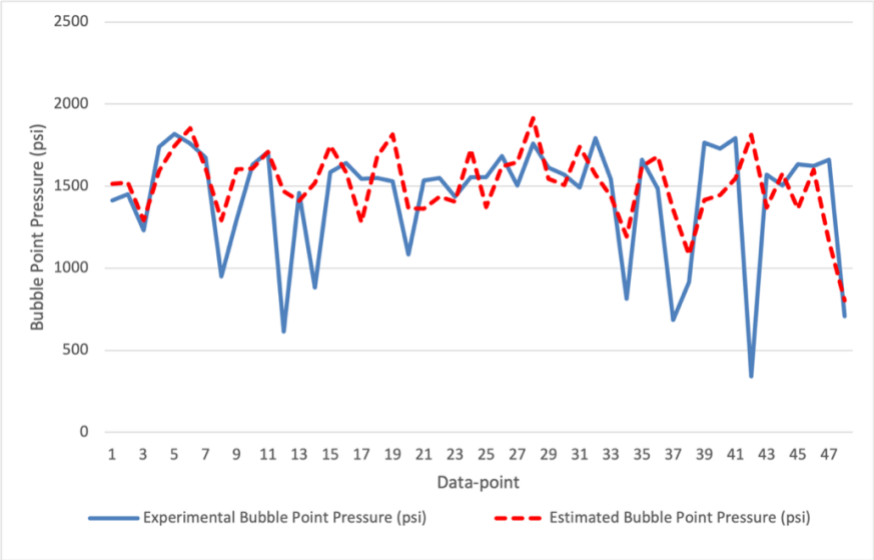

Figure A-13—Cross plot of bubble point pressure for the boosted decision tree regression model (trained on the complete and regional datasets and tested on the incomplete and imputed regional datasets).

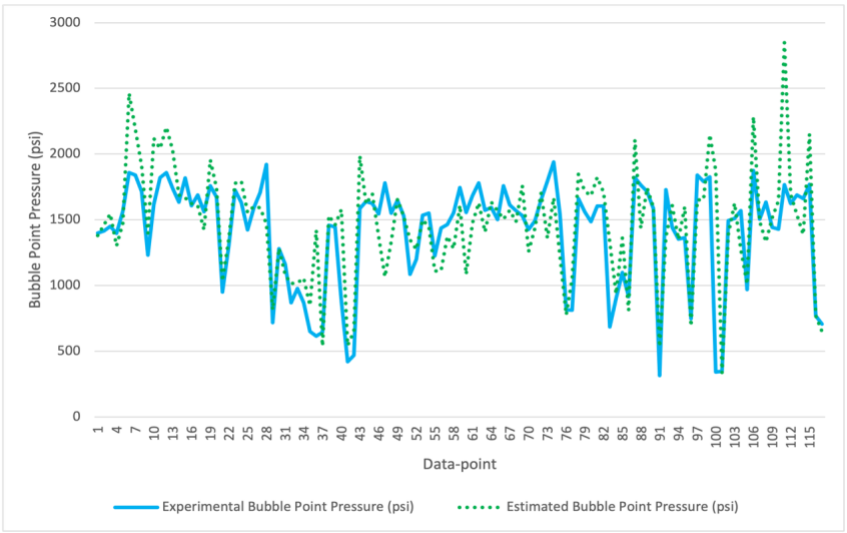

Figure A-14—Cross plot of bubble point pressure for the boosted decision tree regression model (trained on the complete multiregional dataset and tested on the incomplete and imputed regional dataset).

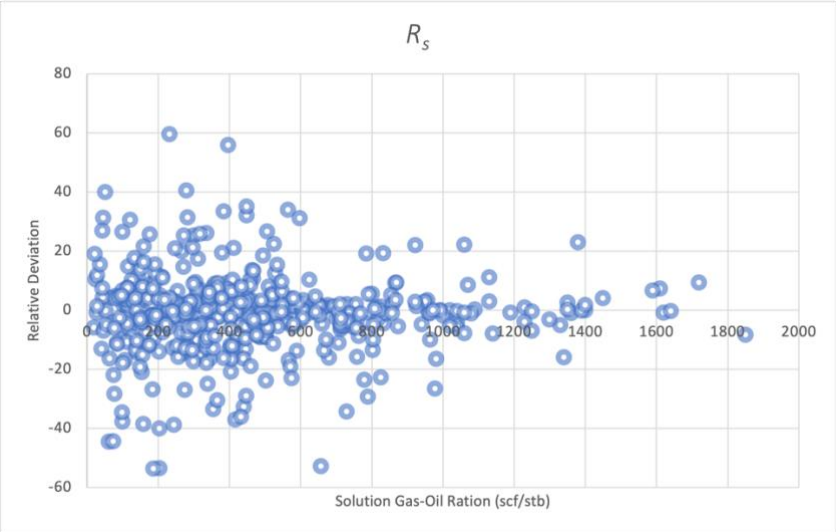

Figure A-15—Cross plot of the relative deviation between the experimental and the estimated values of the bubble point pressure for the boosted decision tree regression model with respect to the input feature R_s (multiregional and complete dataset).

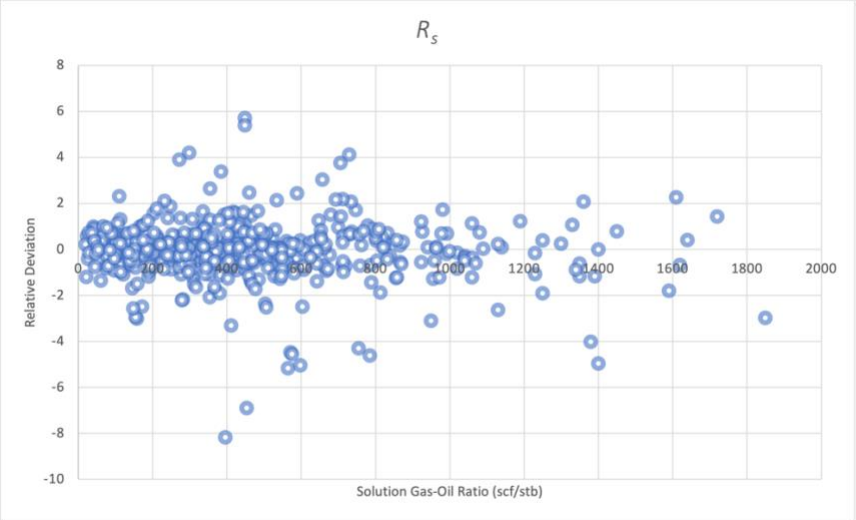

Figure A-16—Cross plot of the relative deviation between the experimental and estimated values of the oil formation volume factor at bubble point pressure for the boosted decision tree regression model with respect to the input feature R_s (multiregional and complete dataset).

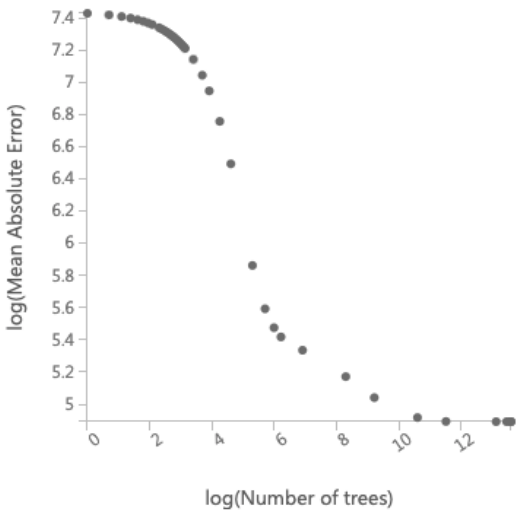

Figure A-17—Cross plot of the mean absolute error against the number of trees constructed by the boosted decision tree regression model for P_b (multiregional and complete dataset).

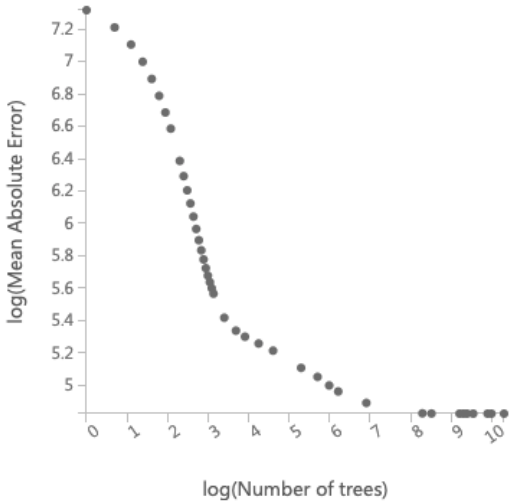

Figure A-18—Cross plot of the mean absolute error against the number of trees constructed by the boosted decision tree regression model for P_b (multiregional and incomplete dataset).

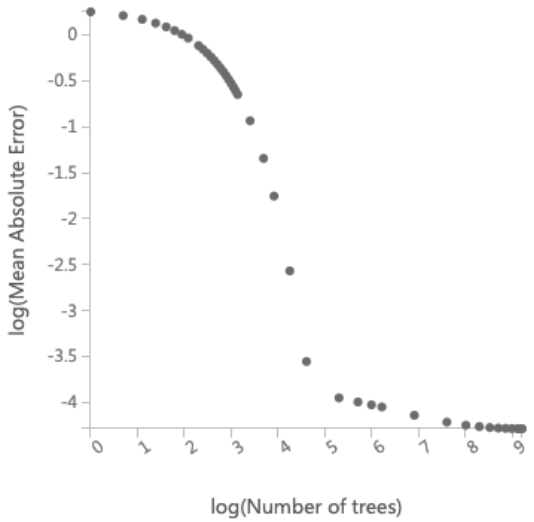

Figure A-19—Cross plot of the mean absolute error against the number of trees constructed by the boosted decision tree regression model for B_{ob} (multiregional and complete dataset).

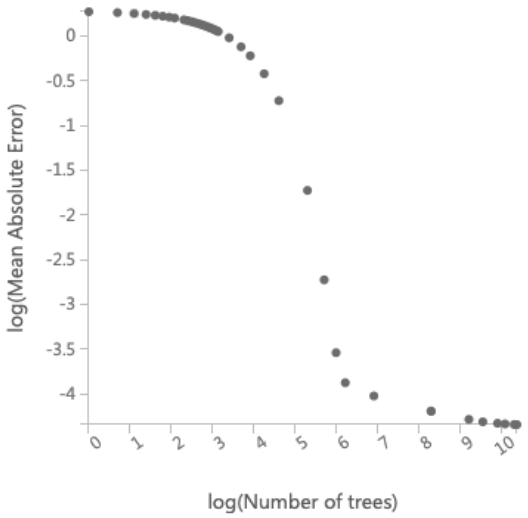

Figure A-20—Cross plot of the mean absolute error against the number of trees constructed by the boosted decision tree regression model for B_{ob} (multiregional and incomplete dataset).

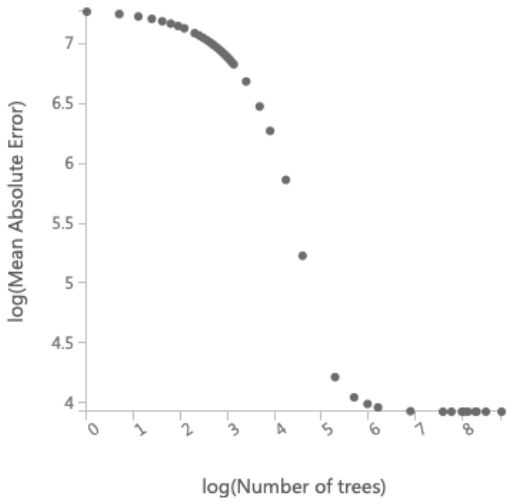

Figure A-21—Cross plot of the mean absolute error against the number of trees constructed by the boosted decision tree regression model for P_b (regional and complete dataset).

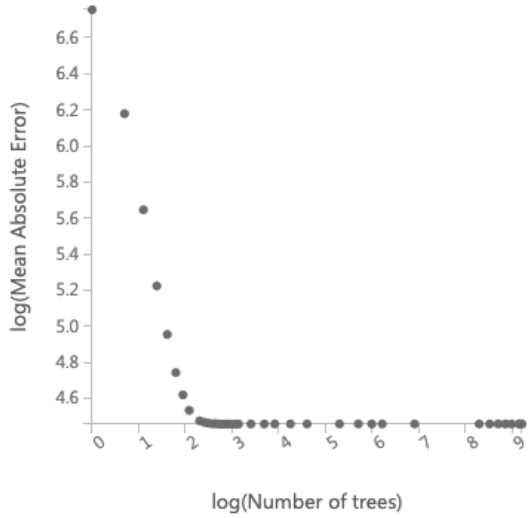

Figure A-22—Cross plot of the mean absolute error against the number of trees constructed by the boosted decision tree regression model for P_b (regional and incomplete dataset).

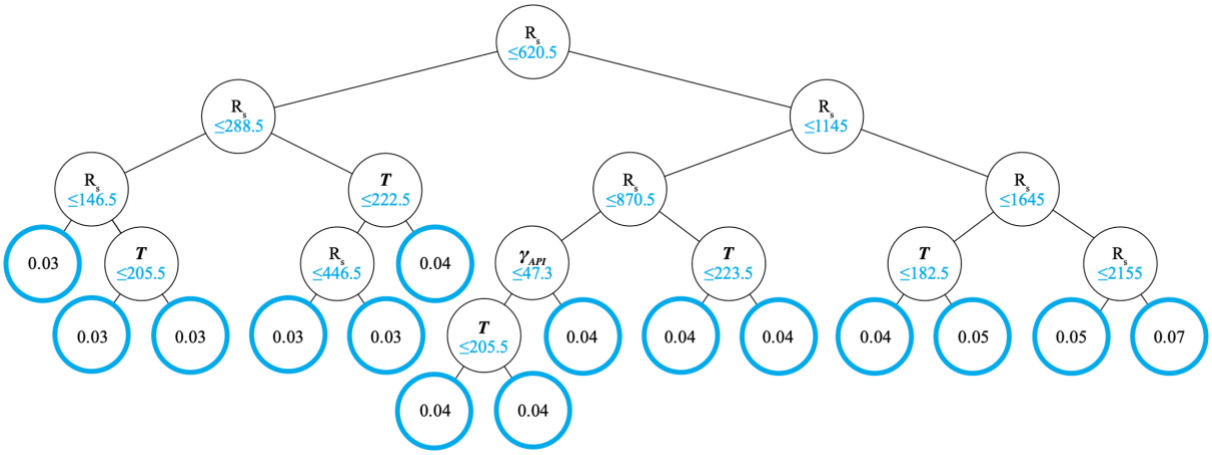

Figure A-23—Tree representation from the gradient boosted decision tree ensemble for B_{ob} (multiregional and complete dataset).

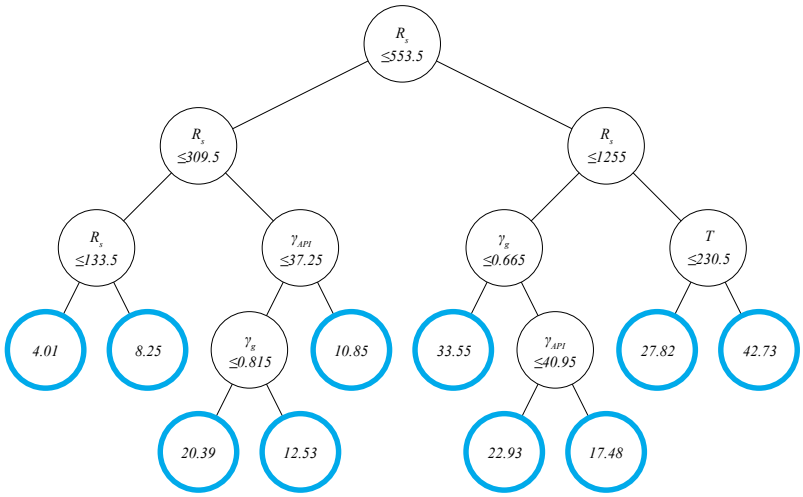

Figure A-24—Tree representation from the gradient boosted decision tree ensemble for P_b (multiregional and complete dataset).

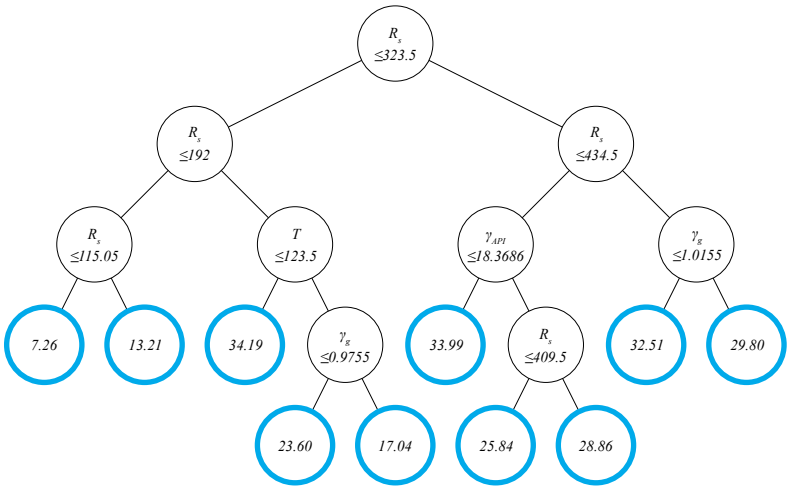

Figure A-25—Tree representation from the gradient boosted decision tree ensemble for P_b (regional and complete dataset).

B. APPENDIX B

The confusion matrices of the DAGs, KNN, decision tree, and random forest models for low- and high-pressure datasets are shown in **Tables B-1–B-26**.

Table B-1—The confusion matrix for the directed acyclic graphs model with the sequential synthetic minority over-sampling technique for the low-pressure datasets (592 kPa).

Class (Flow pattern)	SL	F	DB	SR	A
SL	100%				
F	25%	75%			
DB			100%		
SR	25%			50%	25%
A					100%

Table B-2—The confusion matrix for the directed acyclic graphs model with the synthetic minority over-sampling technique for the low-pressure datasets (592 kPa).

Class (Flow pattern)	SL	F	DB	SR	A
SL	100%				
F	50%	50%			
DB			100%		
SR	25%			50%	25%
A					100%

Table B-3—The confusion matrix for the directed acyclic graphs model with the synthetic minority over-sampling technique plus Tomek for the low-pressure datasets (592 kPa).

Class (Flow pattern)	SL	F	DB	SR	A
SL	100%				
F	50%	50%			
DB			100%		
SR	25%			50%	25%
A					100%

Table B-4—The confusion matrix for the directed acyclic graphs model with the synthetic minority over-sampling technique edited nearest neighbors for the low-pressure datasets (592 kPa).

Class (Flow pattern)	SL	F	DB	SR	A
SL	77.8%	22.2%			
F		100%			
DB			100%		
SR	25%			50%	25%
A					100%

Table B-5—The confusion matrix for the *k*-nearest neighbors model with the synthetic minority over-sampling technique for the low-pressure datasets (592 kPa).

Class (Flow pattern)	SL	F	DB	SR	A
SL	76.9%	7.7%		15.4%	
F		71.4%		28.6%	
DB			100%		
SR	20%			60%	20%
A		50%			50%

Table B-6—The confusion matrix for the decision tree model with the synthetic minority over-sampling technique for the low-pressure datasets (592 kPa).

Class (Flow pattern)	SL	F	DB	SR	A
SL	100%				
F	57.1%	28.6%			14.3%
DB			100%		
SR	20%			80%	
A					100%

Table B-7—The confusion matrix for the random forest model with the synthetic minority over-sampling technique for the low-pressure datasets (592 kPa).

Class (Flow pattern)	SL	F	DB	SR	A
SL	92.3%			7.7%	
F	14.3%	71.4%			14.3%
DB			100%		
SR	20%			40%	40%
A					100%

Table B-8—The confusion matrix for the *k*-nearest neighbors model with the synthetic minority over-sampling technique plus Tomek for the low-pressure datasets (592 kPa).

Class (Flow pattern)	SL	F	DB	SR	A
SL	76.9%	7.7%		15.4%	
F		71.4%		28.6%	
DB			100%		
SR	20%			60%	20%
A		50%			50%

Table B-9—The confusion matrix for the decision tree model with the synthetic minority over-sampling technique plus Tomek for the low-pressure datasets (592 kPa).

Class (Flow pattern)	SL	F	DB	SR	A
SL	100%				
F	42.9%	42.9%			14.2%
DB			100%		
SR	20%			80%	
A					100%

Table B-10—The confusion matrix for the random forest model with the synthetic minority over-sampling technique plus Tomek for the low-pressure datasets (592 kPa).

Class (Flow pattern)	SL	F	DB	SR	A
SL	92.3%			7.7%	
F	14.3%	71.4%			14.3%
DB			100%		
SR	20%			40%	40%
A					100%

Table B-11—The confusion matrix for the *k*-nearest neighbors model with the synthetic minority over-sampling technique edited nearest neighbors for the low-pressure datasets (592 kPa).

Class (Flow pattern)	SL	F	DB	SR	A
SL	69.2%	7.7%	7.7%	15.4%	
F		71.4%		28.6%	
DB			100%		
SR	20%			60%	20%
A		50%			50%

Table B-12—The confusion matrix for the decision tree model with the synthetic minority over-sampling technique edited nearest neighbors for the low-pressure datasets (592 kPa).

Class (Flow Pattern)	SL	F	DB	SR	A
SL	92.3%				7.7%
F		85.7%			14.3%
DB			100%		
SR	20%			80%	
A					100%

Table B-13—The confusion matrix for the random forest model with the synthetic minority over-sampling technique edited nearest neighbors for the low-pressure datasets (592 kPa).

Class (Flow Pattern)	SL	F	DB	SR	A
SL	38.5%	53.8%		7.7%	
F		100%			
DB			100%		
SR	20%			40%	40%
A		50%			50%

Table B-14—The confusion matrix for the directed acyclic graphs model with the sequential synthetic minority over-sampling technique for the high-pressure datasets (2,060 kPa).

Class (Flow pattern)	SL	EB	DB	In-DB	SW-SR	F	SR-A	SS
SL	86.4%	13.6%						
EB		93.3%		6.7%				
DB			100%					
In-DB				100%				
SW-SR	12.5%				87.5%			
F	50%					50%		
SR-A							100%	
SS							50%	50%

Table B-15—The confusion matrix for the directed acyclic graphs model with the synthetic minority over-sampling technique for the high-pressure datasets (2,060 kPa).

Class (Flow pattern)	SL	EB	DB	In-DB	SW-SR	F	SR-A	SS
SL	95.5%	4.5%						
EB		93.3%		6.7%				
DB			100%					
In-DB				100%				
SW-SR	12.5%				87.5%			
F	50%					50%		
SR-A							100%	
SS					50%			50%

Table B-16—The confusion matrix for the directed acyclic graphs model with the synthetic minority over-sampling technique plus Tomek for the of high-pressure datasets (2,060 kPa).

Class (Flow pattern)	SL	EB	DB	In-DB	SW-SR	F	SR-A	SS
SL	90.9%		9.1%					
EB		93.3%		6.7%				
DB			100%					
In-DB				100%				
SW-SR	12.5%				87.5%			
F	50%					50%		
SR-A							100%	
SS					50%			50%

Table B-17—The confusion matrix for the directed acyclic graphs model with the synthetic minority over-sampling technique edited nearest neighbors for the high-pressure datasets (2,060 kPa).

Class (Flow pattern)	SL	EB	DB	In-DB	SW-SR	F	SR-A	SS
SL	72.7%	9.1%		9.1%		9.1%		
EB		93.3%		6.7%				
DB			100%					
In-DB				100%				
SW-SR	12.5%				87.5%			
F					50%	50%		
SR-A							100%	
SS					50%			50%

Table B-18—The confusion matrix for the *k*-nearest neighbors model with the synthetic minority over-sampling technique for the high-pressure datasets (2,060 kPa).

Class (Flow pattern)	SL	EB	DB	In-DB	SW-SR	F	SR-A	SS
SL	55.6%	11.1%		25.9%				7.4%
EB		87.5%		12.5%				
DB	16.7%		83.3%					
In-DB			60%	40%				
SW-SR	27.3%				18.2%	18.2%	9%	27.3%
F					100%			
SR-A					50%		50%	
SS					50%			50%

Table B-19—The confusion matrix for the decision tree model with the synthetic minority over-sampling technique for the high-pressure datasets (2,060 kPa).

Class (Flow pattern)	SL	EB	DB	In-DB	SW-SR	F	SR-A	SS
SL	81.5%	11.1%		3.7%			3.7%	
EB		93.75%		6.25%				
DB			100%					
In-DB				100%				
SW-SR	9.1%				90.9%			
F	50%					50%		
SR-A						50%	50%	
SS								100%

Table B-20—The confusion matrix for the random forest model with the synthetic minority over-sampling technique for the high-pressure datasets (2,060 kPa).

Class (Flow pattern)	SL	EB	DB	In-DB	SW-SR	F	SR-A	SS
SL	59.3%	29.6%		3.7%			7.4%	
EB		93.75%		6.25%				
DB			100%					
In-DB				100%				
SW-SR					81.8%			18.2%
F					25%	50%	25%	
SR-A							100%	
SS					50%			50%

Table B-21—The confusion matrix for the *k*-nearest neighbors model with the synthetic minority over-sampling technique plus Tomek for the high-pressure datasets (2,060 kPa).

Class (Flow pattern)	SL	EB	DB	In-DB	SW-SR	F	SR-A	SS
SL	55.6%	18.5%	7.4%	11.1%				7.4%
EB		75%		25%				
DB	16.7%		83.3%					
In-DB			60%	40%				
SW-SR	27.3%				18.2%	18.2%	9%	27.3%
F					100%			
SR-A					100%			
SS					50%			50%

Table B-22—The confusion matrix for the decision tree model with the synthetic minority over-sampling technique plus Tomek for the high-pressure datasets (2,060 kPa).

Class (Flow pattern)	SL	EB	DB	In-DB	SW-SR	F	SR-A	SS
SL	88.9%	11.1%						
EB		93.75%		6.25%				
DB			100%					
In-DB				100%				
SW-SR	9.1%				81.8%			9.1%
F	25%					75%		
SR-A							100%	
SS					50%			50%

Table B-23—The confusion matrix for the random forest model with the synthetic minority over-sampling technique plus Tomek for the high-pressure datasets (2,060 kPa).

Class (Flow pattern)	SL	EB	DB	In-DB	SW-SR	F	SR-A	SS
SL	59.3%	29.6%		3.7%			7.4%	
EB		100%						
DB			100%					
In-DB				100%				
SW-SR					81.8%			18.2%
F					25%	25%	50%	
SR-A							100%	
SS					50%			50%

Table B-24—The confusion matrix for the *k*-nearest neighbors model with the synthetic minority over-sampling technique edited nearest neighbors for the high-pressure datasets (2,060 kPa).

Class (Flow pattern)	SL	EB	DB	In-DB	SW-SR	F	SR-A	SS
SL	55.6%	11.1%	7.4%	18.5%				7.4%
EB		68.75%		31.25%				
DB	16.7%		83.3%					
In-DB			40%	60%				
SW-SR	27.3%				18.2%	18.2%	9%	27.3%
F					100%			
SR-A					100%			
SS					50%			50%

Table B-25—The confusion matrix for the decision tree model with the synthetic minority over-sampling technique edited nearest neighbors for the high-pressure datasets (2,060 kPa).

Class (Flow pattern)	SL	EB	DB	In-DB	SW-SR	F	SR-A	SS
SL	74.1%	11.1%		7.4%			7.4%	
EB		93.75%		6.25%				
DB			100%					
In-DB				100%				
SW-SR	9.1%				72.7%			18.2%
F	25%					75%		
SR-A							100%	
SS								100%

Table B-26—The confusion matrix for the random forest model with the synthetic minority over-sampling technique edited nearest neighbors for the high-pressure datasets (2,060 kPa).

Class (Flow pattern)	SL	EB	DB	In-DB	SW-SR	F	SR-A	SS
SL	48.2%	29.6%		14.8%		7.4%		
EB		100%						
DB			100%					
In-DB				100%				
SW-SR		9.1%			72.7%			18.2%
F					25%	75%		
SR-A							100%	
SS								100%

C. APPENDIX C

The classification reports for the trained KNN, decision tree, and random forest models with SMOTE, SMOTETomek, and SMOTEENN low- and high-pressure datasets are shown in **Tables C-1–C-18**.

Table C-1—Classification report for the trained *k*-nearest neighbors model with synthetic minority over-sampling technique low-pressure datasets.

	Precision	*Recall*	*F_1-score*	Support
A	0.50	0.50	0.50	2
DB	1.00	1.00	1.00	5
F	0.71	0.71	0.71	7
SL	0.91	0.77	0.83	13
SR	0.43	0.60	0.50	5
Accuracy			0.75	32
Macro avg	0.71	0.72	0.71	32
Weighted avg	0.78	0.75	0.76	32

Table C-2—Classification report for the trained decision tree model with synthetic minority over-sampling technique low-pressure datasets.

	Precision	*Recall*	*F_1-score*	Support
A	0.67	1.00	0.80	2
DB	1.00	1.00	1.00	5
F	1.00	0.29	0.44	7
SL	0.72	1.00	0.84	13
SR	1.00	0.80	0.89	5
Accuracy			0.81	32
Macro avg	0.88	0.82	0.79	32
Weighted avg	0.87	0.81	0.78	32

Table C-3—Classification report for the trained random forest model with synthetic minority over-sampling technique low-pressure datasets.

	Precision	*Recall*	*F_1-score*	Support
A	0.40	1.00	0.57	2
DB	1.00	1.00	1.00	5
F	1.00	0.71	0.83	7
SL	0.86	0.92	0.89	13
SR	0.67	0.40	0.50	5
Accuracy			0.81	32
Macro avg	0.78	0.81	0.76	32
Weighted avg	0.85	0.81	0.81	32

Table C-4—Classification report for the trained *k*-nearest neighbors model with synthetic minority over-sampling technique plus Tomek low-pressure datasets.

	Precision	*Recall*	*F_1-score*	Support
A	0.50	0.50	0.50	2
DB	1.00	1.00	1.00	5
F	0.71	0.71	0.71	7
SL	0.91	0.77	0.83	13
SR	0.43	0.60	0.50	5
Accuracy			0.75	32
Macro avg	0.71	0.72	0.71	32
Weighted avg	0.78	0.75	0.76	32

Table C-5—Classification report for the trained decision tree model with synthetic minority over-sampling technique plus Tomek low-pressure datasets.

	Precision	*Recall*	*F_1-score*	Support
A	0.67	1.00	0.80	2
DB	1.00	1.00	1.00	5
F	1.00	0.43	0.60	7
SL	0.76	1.00	0.87	13
SR	1.00	0.80	0.89	5
Accuracy			0.84	32
Macro avg	0.89	0.85	0.83	32
Weighted avg	0.88	0.84	0.83	32

Table C-6—Classification report for the trained random forest model with synthetic minority over-sampling technique plus Tomek low-pressure datasets.

	Precision	*Recall*	*F_1-score*	Support
A	0.40	1.00	0.57	2
DB	1.00	1.00	1.00	5
F	1.00	0.71	0.83	7
SL	0.86	0.92	0.89	13
SR	0.67	0.40	0.50	5
Accuracy			0.81	32
Macro avg	0.78	0.81	0.76	32
Weighted avg	0.85	0.81	0.81	32

Table C-7—Classification report for the trained *k*-nearest neighbors model with synthetic minority over-sampling technique edited nearest neighbors low-pressure datasets.

	Precision	*Recall*	*F_1-score*	Support
A	0.50	0.50	0.50	2
DB	0.83	1.00	0.91	5
F	0.71	0.71	0.71	7
SL	0.90	0.69	0.78	13
SR	0.43	0.60	0.50	5
Accuracy			0.72	32
Macro avg	0.68	0.70	0.68	32
Weighted avg	0.75	0.72	0.73	32

Table C-8—Classification report for the trained decision tree model with synthetic minority over-sampling technique edited nearest neighbors low-pressure datasets.

	Precision	*Recall*	*F_1-score*	Support
A	0.50	1.00	0.67	2
DB	1.00	1.00	1.00	5
F	1.00	0.86	0.92	7
SL	0.92	0.92	0.92	13
SR	1.00	0.80	0.89	5
Accuracy			0.91	32
Macro avg	0.88	0.92	0.88	32
Weighted avg	0.94	0.91	0.91	32

Table C-9—Classification report for the trained random forest model with synthetic minority over-sampling technique edited nearest neighbors low-pressure datasets.

	Precision	*Recall*	*F_1-score*	Support
A	0.33	0.50	0.40	2
DB	1.00	1.00	1.00	5
F	0.47	1.00	0.64	7
SL	0.83	0.38	0.53	13
SR	0.67	0.40	0.50	5
Accuracy			0.62	32
Macro avg	0.66	0.66	0.61	32
Weighted avg	0.72	0.62	0.61	32

Table C-10—Classification report for the trained *k*-nearest neighbors model with synthetic minority over-sampling technique high-pressure datasets.

	Precision	*Recall*	*F_1-score*	Support
DB	0.62	0.83	0.71	6
EB	0.82	0.88	0.85	16
F	0.00	0.00	0.00	4
In-DB	0.18	0.40	0.25	5
SL	0.79	0.56	0.65	27
SR-A	0.50	0.50	0.50	2
SS	0.17	0.50	0.25	2
SW-SR	0.25	0.18	0.21	11
Accuracy			0.55	73
Macro avg	0.42	0.48	0.43	73
Weighted avg	0.59	0.55	0.56	73

Table C-11—Classification report for the trained decision tree model with synthetic minority over-sampling technique high-pressure datasets.

	Precision	*Recall*	*F_1-score*	Support
DB	1.00	1.00	1.00	6
EB	0.83	0.94	0.88	16
F	0.67	0.50	0.57	4
In-DB	0.71	1.00	0.83	5
SL	0.88	0.81	0.85	27
SR-A	0.50	0.50	0.50	2
SS	1.00	1.00	1.00	2
SW-SR	1.00	0.91	0.95	11
Accuracy			0.86	73
Macro avg	0.82	0.83	0.82	73
Weighted avg	0.87	0.86	0.86	73

Table C-12—Classification report for the trained random forest model with synthetic minority over-sampling technique high-pressure datasets.

	Precision	*Recall*	*F_1-score*	Support
DB	1.00	1.00	1.00	6
EB	0.65	0.94	0.77	16
F	1.00	0.50	0.67	4
In-DB	0.71	1.00	0.83	5
SL	1.00	0.59	0.74	27
SR-A	0.40	1.00	0.57	2
SS	0.33	0.50	0.40	2
SW-SR	0.82	0.82	0.82	11
Accuracy			0.77	73
Macro avg	0.74	0.79	0.73	73
Weighted avg	0.84	0.77	0.77	73

Table C-13—Classification report for the trained *k*-nearest neighbors model with synthetic minority over-sampling technique plus Tomek high-pressure datasets.

	Precision	*Recall*	*F_1-score*	Support
DB	0.50	0.83	0.62	6
EB	0.71	0.75	0.73	16
F	0.00	0.00	0.00	4
In-DB	0.22	0.40	0.29	5
SL	0.79	0.56	0.65	27
SR-A	0.00	0.00	0.00	2
SS	0.17	0.50	0.25	2
SW-SR	0.22	0.18	0.20	11
Accuracy			0.51	73
Macro avg	0.33	0.40	0.34	73
Weighted avg	0.54	0.51	0.51	73

Table C-14—Classification report for the trained decision tree model with synthetic minority over-sampling technique plus Tomek high-pressure datasets.

	Precision	*Recall*	*F_1-score*	Support
DB	1.00	1.00	1.00	6
EB	0.83	0.94	0.88	16
F	1.00	0.75	0.86	4
In-DB	0.83	1.00	0.91	5
SL	0.92	0.89	0.91	27
SR-A	1.00	1.00	1.00	2
SS	0.50	0.50	0.50	2
SW-SR	0.90	0.82	0.86	11
Accuracy			0.89	73
Macro avg	0.87	0.86	0.86	73
Weighted avg	0.89	0.89	0.89	73

Table C-15—Classification report for the trained random forest model with synthetic minority over-sampling technique plus Tomek high-pressure datasets.

	Precision	*Recall*	*F_1-score*	Support
DB	1.00	1.00	1.00	6
EB	0.67	1.00	0.80	16
F	1.00	0.25	0.40	4
In-DB	0.83	1.00	0.91	5
SL	1.00	0.59	0.74	27
SR-A	0.33	1.00	0.50	2
SS	0.33	0.50	0.40	2
SW-SR	0.82	0.82	0.82	11
Accuracy			0.77	73
Macro avg	0.75	0.77	0.70	73
Weighted avg	0.85	0.77	0.76	73

Table C-16—Classification report for the trained *k*-nearest neighbors model with synthetic minority over-sampling technique edited nearest neighbors high-pressure datasets.

	Precision	*Recall*	*F_1-score*	Support
DB	0.56	0.83	0.67	6
EB	0.79	0.69	0.73	16
F	0.00	0.00	0.00	4
In-DB	0.23	0.60	0.33	5
SL	0.79	0.56	0.65	27
SR-A	0.00	0.00	0.00	2
SS	0.17	0.50	0.25	2
SW-SR	0.22	0.18	0.20	11
Accuracy			0.51	73
Macro avg	0.34	0.42	0.35	73
Weighted avg	0.56	0.51	0.52	73

Table C-17—Classification report for the trained decision tree model with synthetic minority over-sampling technique edited nearest neighbors high-pressure datasets.

	Precision	*Recall*	*F_1-score*	Support
DB	1.00	1.00	1.00	6
EB	0.83	0.94	0.88	16
F	1.00	0.75	0.86	4
In-DB	0.62	1.00	0.77	5
SL	0.91	0.74	0.82	27
SR-A	0.50	1.00	0.67	2
SS	0.50	1.00	0.67	2
SW-SR	1.00	0.73	0.84	11
Accuracy			0.84	73
Macro avg	0.80	0.89	0.81	73
Weighted avg	0.88	0.84	0.84	73

Table C-18—Classification report for the trained random forest model with synthetic minority over-sampling technique edited nearest neighbors high-pressure datasets.

	Precision	*Recall*	*F_1-score*	Support
DB	1.00	1.00	1.00	6
EB	0.64	1.00	0.78	16
F	0.60	0.75	0.67	4
In-DB	0.56	1.00	0.71	5
SL	1.00	0.48	0.65	27
SR-A	1.00	1.00	1.00	2
SS	0.50	1.00	0.67	2
SW-SR	0.89	0.73	0.80	11
Accuracy			0.75	73
Macro avg	0.77	0.87	0.78	73
Weighted avg	0.84	0.75	0.75	73

D. APPENDIX D

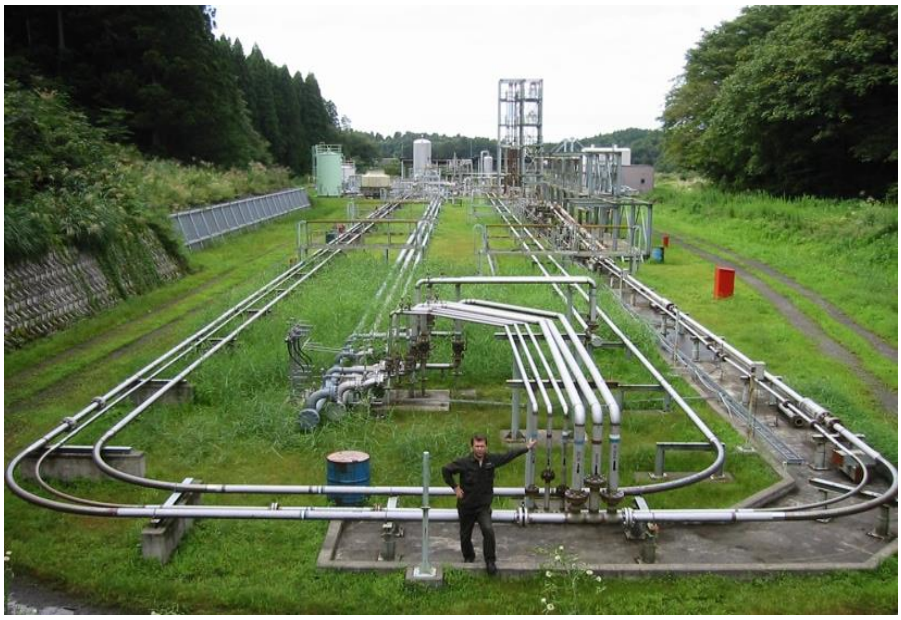

Figure D-1—Kashiwazaki two-phase flow-test facility.

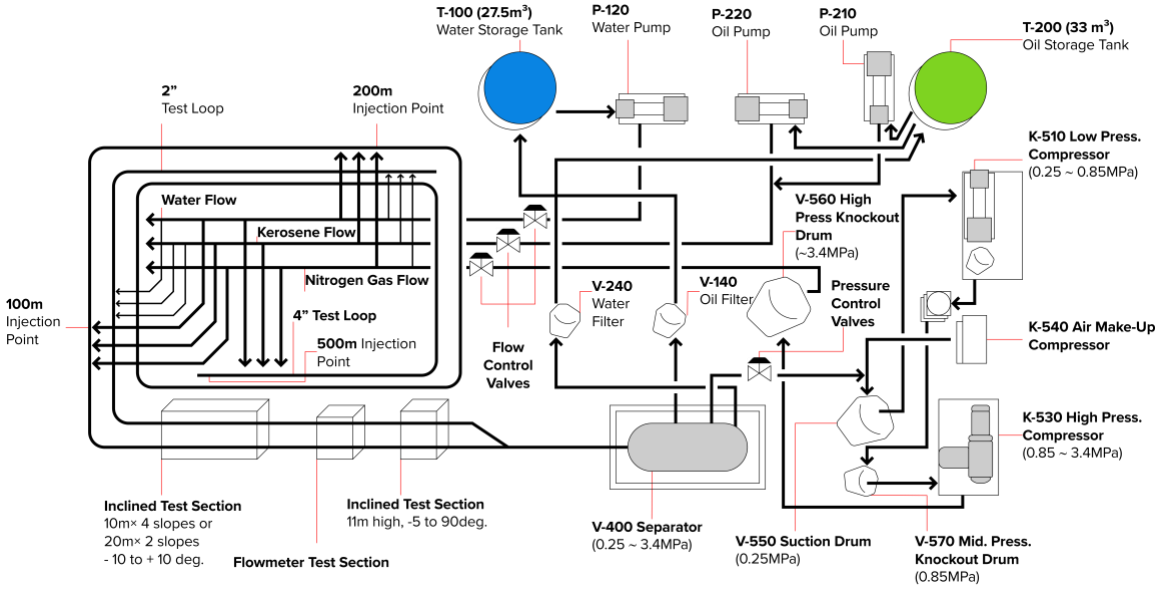

Figure D-2—Schematic diagram of the test facility.

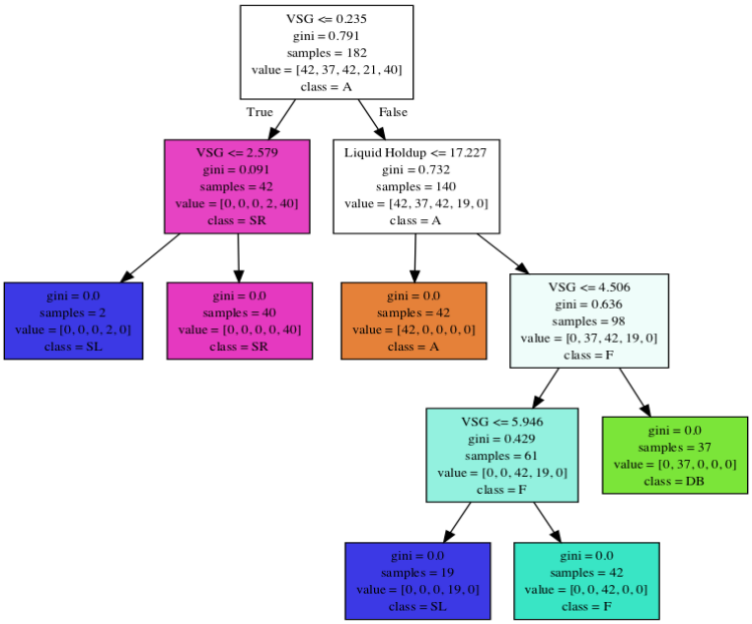

Figure D-3—Decision tree classifier for the low-pressure dataset with synthetic minority over-sampling technique edited nearest neighbors preprocessing technique.

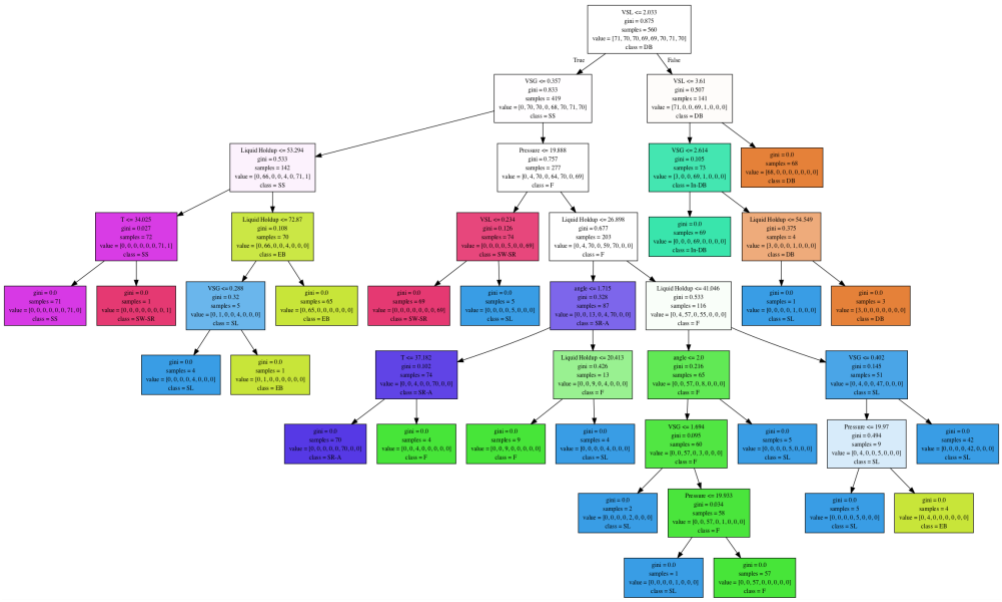

Figure D-4—Decision tree classifier for the high-pressure dataset with synthetic minority over-sampling technique plus Tomek preprocessing technique.

E. APPENDIX E

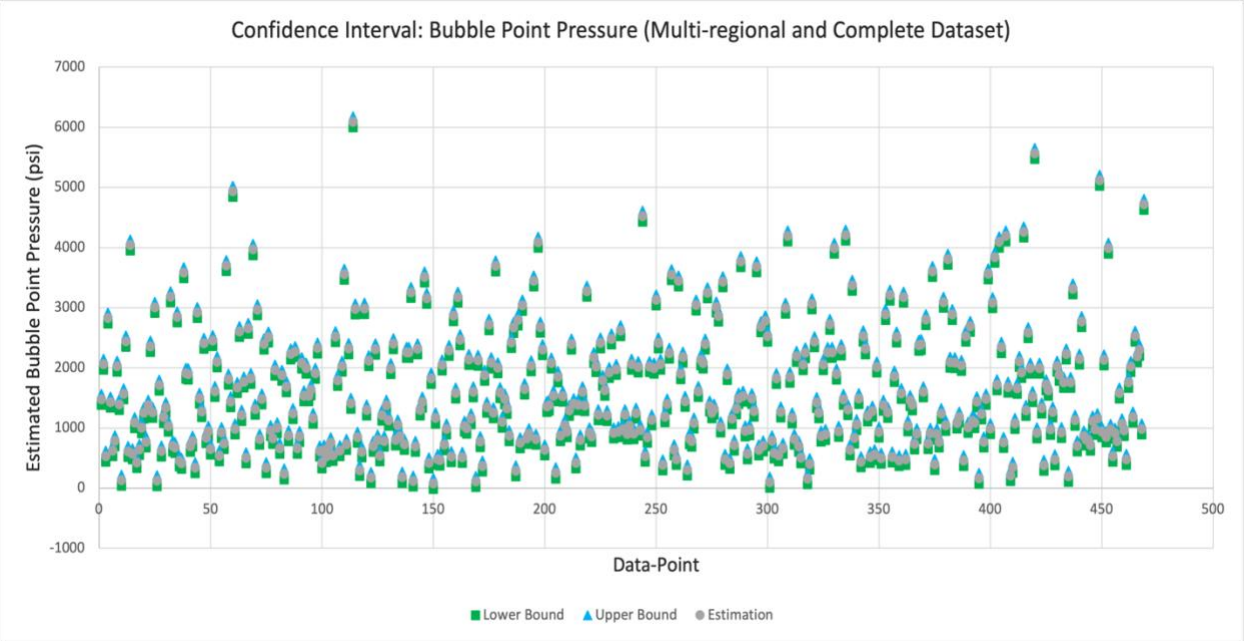

Figure E-1—Confidence interval for P_b (multi-regional and complete dataset).

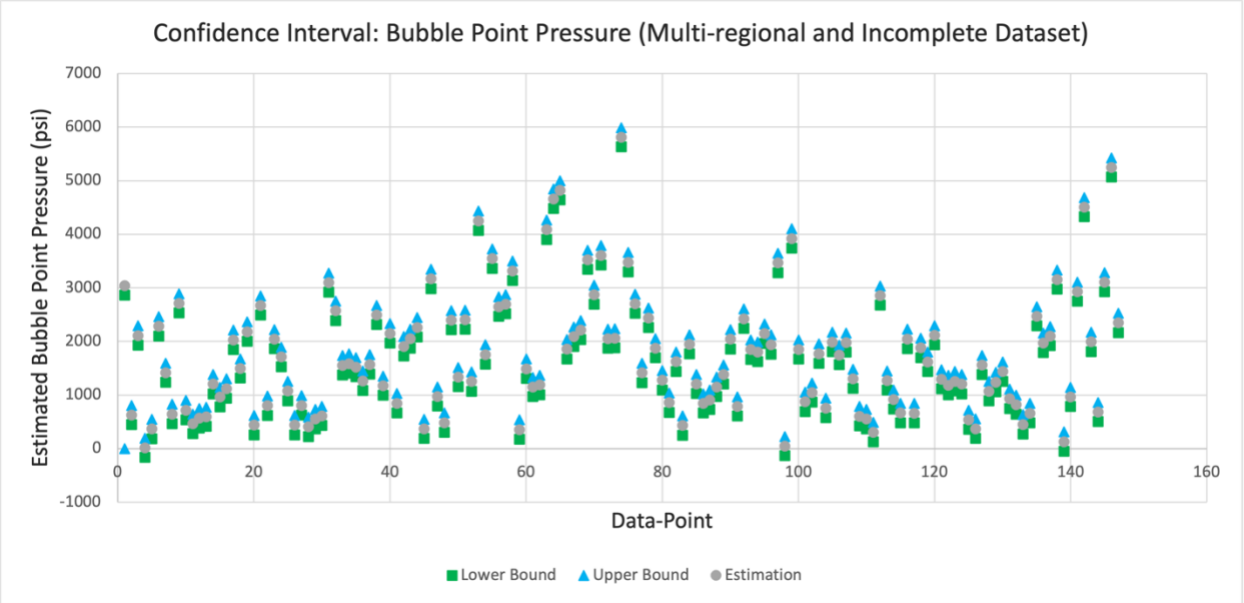

Figure E-2—Confidence interval for P_b (multi-regional and incomplete dataset).

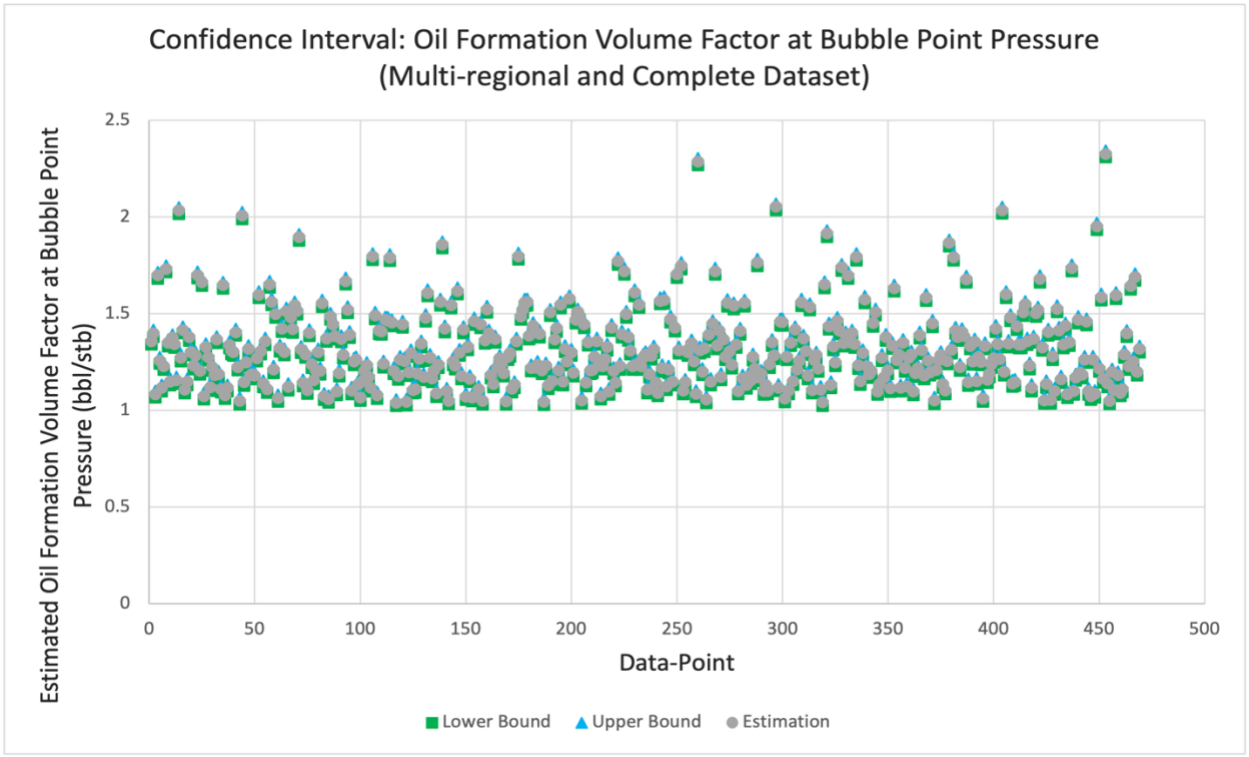

Figure E-3—Confidence interval for B_{ob} (multi-regional and complete dataset).

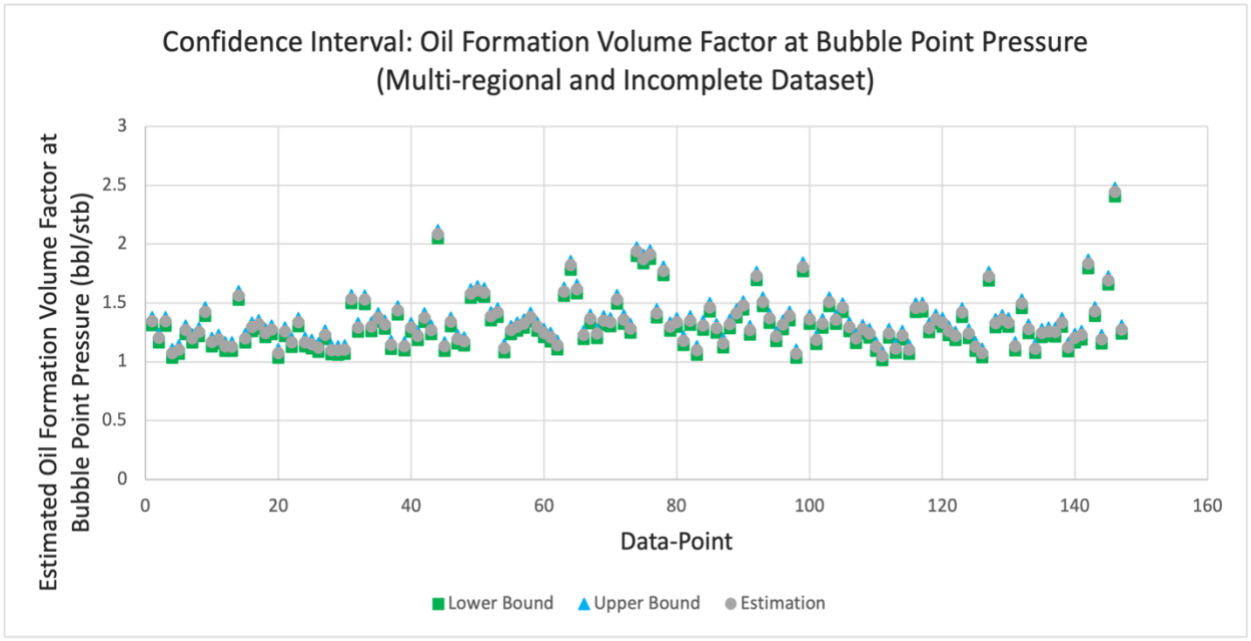

Figure E-4—Confidence interval for B_{ob} (multi-regional and incomplete dataset).

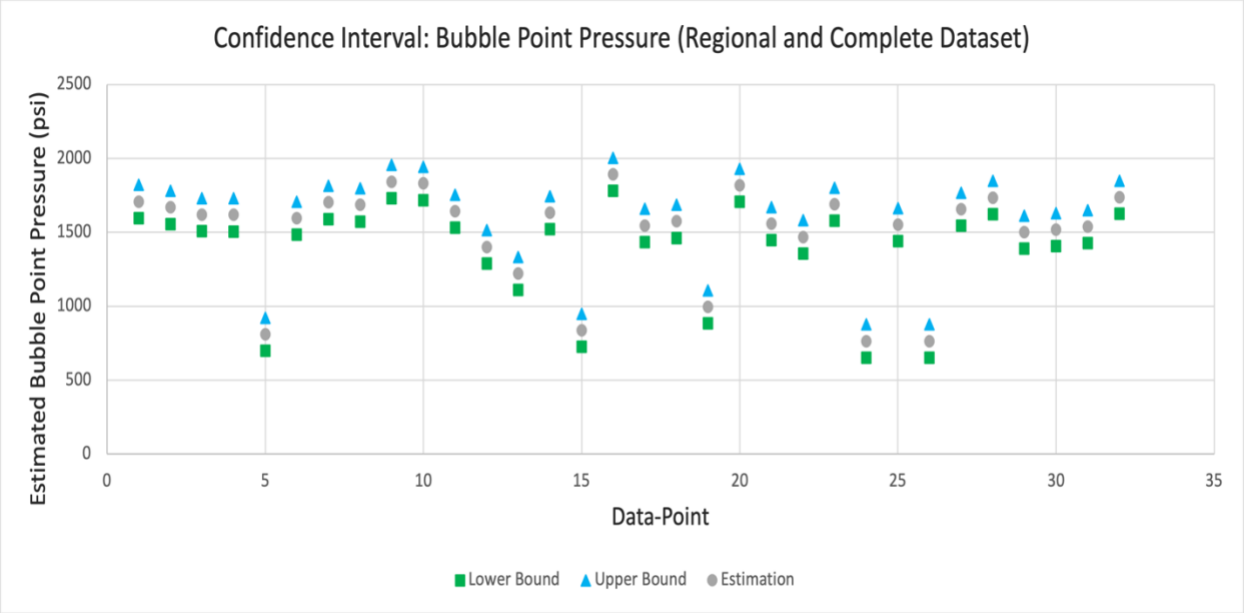

Figure E-5—Confidence interval for P_b (regional and complete dataset).

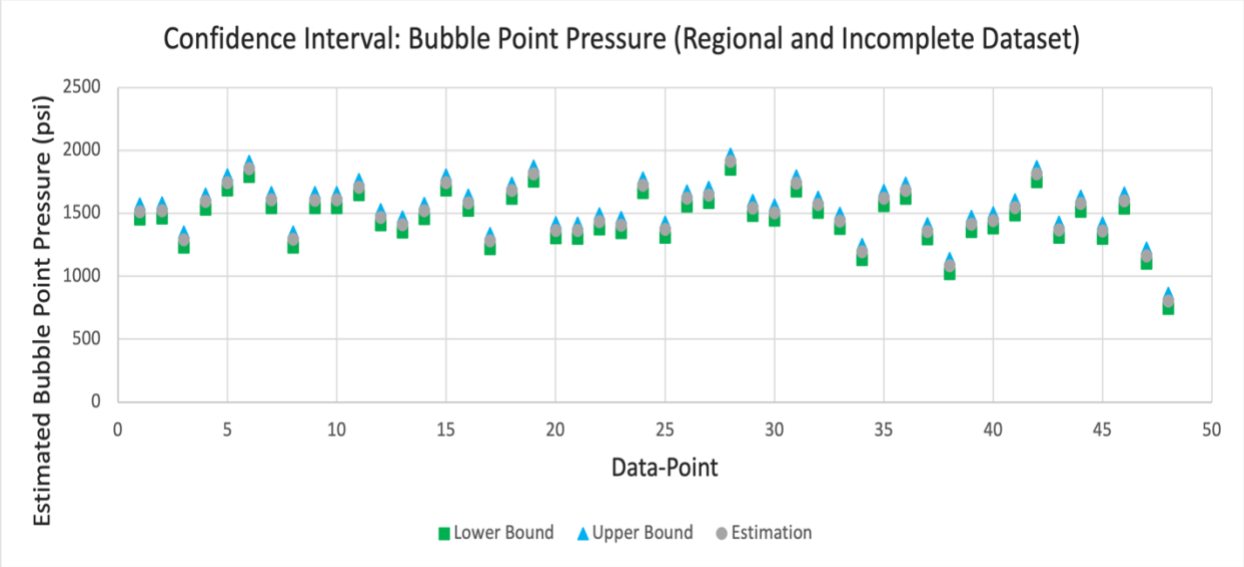

Figure E-6—Confidence interval for P_b (regional and incomplete dataset).

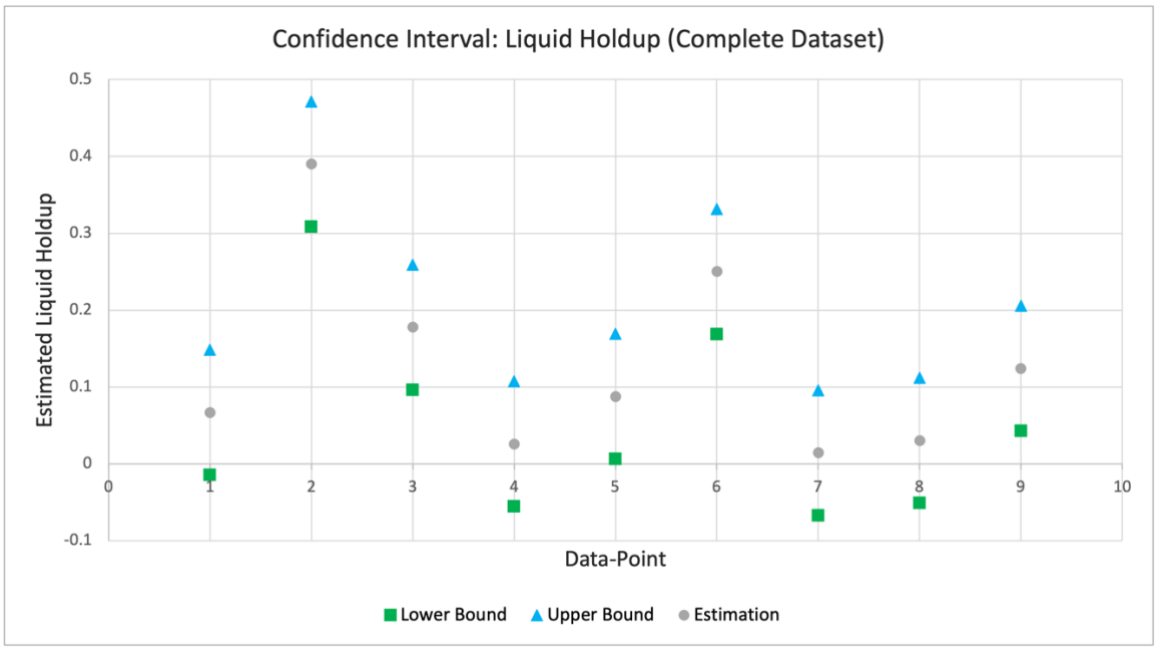

Figure E-7—Confidence interval for H_L (complete dataset).

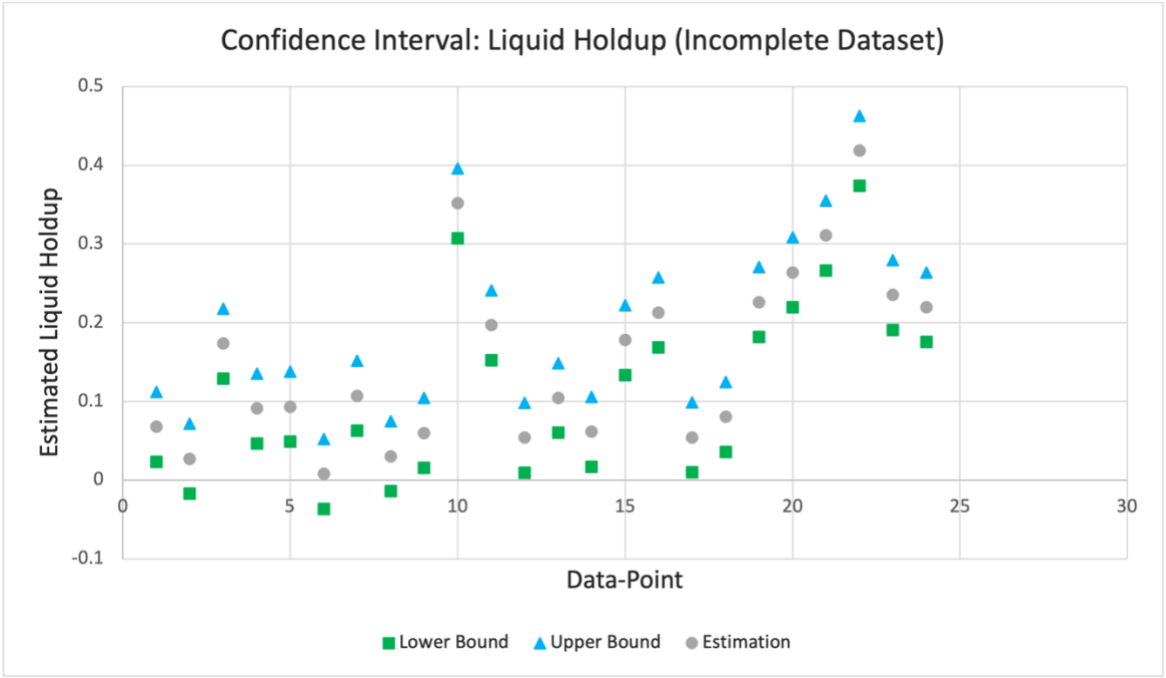

Figure E-8—Confidence interval for H_L (incomplete dataset).

F. APPENDIX F

The results of the cross-validation with complete datasets, and incomplete data in the training, validation, and testing datasets, are shown in **Table F-1** and **Table F-2**. The hyperparameter settings are listed in Table A-4 and Table A-6.

For the incomplete multiregional dataset, for each input feature there are 200 missing data points (4% of the dataset) with imputation. For the incomplete regional dataset, for each input feature there are 19 missing data points (16% of the dataset) with imputation. For the incomplete liquid holdup dataset, for each input feature there are 11 missing data points (10% of the dataset) with imputation.

Table F-1—Cross-validation results for P_b and B_{ob}.

Dataset	Number of folds	Number of samples per fold	E_a
P_b multi-regional (complete)	10	520~521	8.9
P_b multi-regional (incomplete)	10	520~521	12.04
B_{ob} multi-regional (complete)	10	520~521	0.86
B_{ob} multi-regional (incomplete)	10	520~521	1.04
P_b regional (complete)	10	11~12	6.71
P_b regional (incomplete)	10	11~12	19.56

Table F-2—Cross-validation results for H_L.

Dataset	Number of folds	Number of samples per fold	E_a
Liquid holdup (complete)	10	11~12	13.97
Liquid holdup (incomplete)	10	11~12	46.99