RGNZ/MGPB

Flow Pattern Prediction in Horizontal and Inclined Pipes using Tree-Based Automated Machine Learning

Rudarsko-geološko-naftni zbornik
(The Mining-Geology-Petroleum Engineering Bulletin)
UDC: 620.1064: 006.31
DOI: 10.17794/rgn.2024.4.12

Original scientific paper

Agash Uthayasuriyan[1]; Ugochukwu Ilozurike Duru[2]; Angela Nwachukwu[3]; Thangavelu Shunmugasundaram[4]; Jeyakumar Gurusamy[5]

[1] *Department of Multidisciplinary Graduate Engineering, Northeastern University, Boston, United States of America, orcid.org/0009-0006-1350-7590*

[2] *Department of Petroleum Engineering, Federal University of Technology, Owerri, Nigeria, orcid.org/ 0000-0001-7920-8047*

[3] *Department of Petroleum Engineering, Federal University of Technology, Owerri, Nigeria, orcid.org/ 0000-0003-1340-439X*

[4] *Department of Computer Science and Engineering, Amrita School of Computing, Coimbatore, India, orcid.org/0000-0001-6505-4016*

[5] *Department of Computer Science and Engineering, Amrita School of Computing, Coimbatore, India, orcid.org/0000-0002-5501-2338*

Abstract

In the oil and gas industry, understanding two-phase (gas-liquid) flow is pivotal, as it directly influences equipment design, quality control, and operational efficiency. Flow pattern determination is thus fundamental to industrial engineering and management. This study utilizes the Tree-based Pipeline Optimization Tool (TPOT), an Automated Machine Learning (AutoML) framework that employs genetic programming, in obtaining the best machine learning model for a provided dataset. This paper presents the design of flow pattern prediction models using the TPOT. The TPOT was applied to predict flow patterns in 2.5 cm and 5.1 cm diameter pipes, using datasets from existing literature. The datasets went through handling of imbalanced data, standardization, and one-hot encoding as data preparation techniques before being fed into TPOT. The models designed for the 2.5 cm and 5.1 cm datasets were named as FPTL_TPOT_2.5 and FPTL_TPOT_5.1, respectively. A comparative analysis of these models alongside other standard supervised machine learning models and similar state-of-the-art similar two-phase flow prediction models was carried out and the insights on the performance of these TPOT designed models were discussed. The results demonstrated that models designed with TPOT achieve remarkable accuracy, scoring 97.66% and 98.09%, for the 2.5 cm and 5.1 cm datasets respectively. Furthermore, the FPTL_TPOT_2.5 and FPTL_TPOT_5.1 models outperformed other counterpart machine learning models in terms of performance, underscoring TPOT's effectiveness in designing machine learning models for flow pattern prediction. The findings of this research carry significant implications for enhancing efficiency and optimizing industrial processes in the oil and gas sector.

Keywords:
multiphase flow; two-phase flow; flow pattern prediction; machine learning; AutoML

1. Introduction

Some engineering industries, such as chemical, geothermal, and petroleum, experience multiphase flow (**Abduvayt et al., 2003; Shoham, 2006; Jahanandish et al., 2011, Malbrel et al., 2024**). Multiphase flow could be liquid-liquid, gas-liquid, or even solid-liquid in the case of two-phase flow. In the case of three-phase flows it could be gas-oil-water, oil-water-solid or gas-liquid-solid. Two-phase gas-liquid is the most common multiphase flow in the oil and gas industry and is considered as an important research domain. The uniqueness and the properties of phase components, length and diameter of well-bore tubing, pipeline, and inclination angle contribute to the complexity of multiphase flow in the petroleum industry. Prediction of multiphase flow characteristics (flow pattern and liquid hold-up) in the petroleum industry is key for pressure gradient determination, which is very necessary for sizing production facilities in the field for optimum production (**Attia et al., 2013; Duru et al., 2022**). Researchers have carried out studies to achieve optimum production through various methods of pressure gradient prediction. Flow pattern is a spatial distribution of different phases of fluid flowing simultaneously in a pipe or conduit. It is an important characteristic of multiphase flow in several industrial applications (**Lin et al., 2020**) and several flow patterns have been identified in horizontal and vertical flow in the industry.

On the other hand, Machine Learning is an active domain under Artificial Intelligence (AI), focusing on designing systems or models for making computer systems to learn, predict and make decisions based on the given

Corresponding author: Jeyakumar Gurusamy
e-mail address: g_jeyakumar@cb.amrita.edu

data, without being explicitly programmed. It involves the design and development of computational models that can analyse and interpret large amounts of data, identify patterns, and learn from them to improve performance over time. Although the machine learning models provide an ability to predict the output from a given dataset, the accuracy of the machine learning model is considered extremely important.

Since the accuracy of the model directly relates to the precision of the prediction made by the machine learning model, it is extremely important to design a machine learning model that has a higher level of accuracy without over-fitting the data. To achieve this, several approaches can be employed, including the utilization of various models (**Hernandez et al., 2019; Rushd et al., 2022; Mask et al., 2019**), hyperparameter tuning for optimizing the machine learning model (**Uthayasuriyan et al., 2023; Muthaiah et al., 2019; Batchu and Seetha, 2021**), and hybridization of the machine learning model with Evolutionary Algorithms (EAs) (**Jayakumar and Raju, 2011; Anusha et al., 2015, Uthayasuriyan et al., 2024**). However, it is a time-consuming process to evaluate all the designed models in terms of achieved accuracy.

With the help of AutoML, the process of building and optimizing the Machine Learning pipelines can be done effectively (**Spandana et al., 2023**). Among the available AutoML models, Tree-based Pipeline Optimization Tool (TPOT) follows the work flow of genetic programming algorithm to search and select the best machine learning model for a given dataset. It explores a large search space of possible machine learning model pipelines, including data preprocessing, feature selection and dimensionality reduction techniques. This paper aims to design machine learning models using the TPOT library for predicting the flow pattern in two-phase gas-liquid flows and provides a comparison of the resulting models with other standard machine learning models.

2. Automated Machine Learning with TPOT

The core concept of machine learning revolves around creating mathematical models that can automatically learn and adapt by learning the patterns in data or through previous experiences (**Hafsa et al., 2023; Uthayasuriyan et al., 2023**). These models can be trained using a diverse range of datatypes, such as images, text, or numerical values, and various techniques are employed to extract meaningful features and relationships from the data. The machine learning algorithms utilize these extracted features to generalize and make predictions or take actions on new, unseen data. One of the key advantages of machine learning is its ability to handle complex problems and large datasets more efficiently than traditional rule-based programming (**Kim et al., 2020; Manami et al., 2023; Barjouei et al., 2021**).

The presence of numerous machine learning algorithms and the complexity involved in tuning its hyperparameters makes it difficult to evaluate them and to find the most suitable model for a particular dataset. To ease this process, AutoML (Automated Machine Learning) has been used. AutoML utilizes advanced algorithms, optimization techniques, and heuristics to automatically search, evaluate, and select the best performing machine learning models.

TPOT is an AutoML library that has genetic programming in its framework (**Le et al., 2020, Olson et al., 2016a, Olson et al., 2016b**). This allows TPOT to automatically discover complex combinations of pre-processing and modelling steps that might be challenging to determine manually. It is capable of finding the right machine learning model to perform classification, regression, and time series forecasting tasks (**Yusof et al., 2024**). The working of TPOT involves several steps that include initialization, evaluation, followed by genetic programming and producing the output of the best-found machine learning model pipelines. The workflow of TPOT is represented in **Figure 1**.

TPOT randomly generates several pipelines as an initial population, where each pipeline consists of a random sequence of machine-learning operators. After initialization, the fitness of each of the pipelines is evaluated using metrics such as accuracy, and mean squared error. Additionally, cross-validation is done to assess the pipeline's performance on different subsets of the training data. The fitness score reflects how well the pipeline performs on the evaluation metric.

Genetic programming, as depicted in **Figure 2**, is used to evolve the population of pipelines over multiple generations. It selects the best-performing pipelines based on their fitness scores, that are put in the mating pool. The pipelines present in the mating pool undergo the genetic operators of crossover and mutation to create new pipeline offspring. In the crossover operation, TPOT selects two parent pipelines and combines their sequence of operators to create a new offspring pipeline. The combination can occur at a specific operator or a subsequence of operators. Whereas in the mutation operation, TPOT randomly modifies an operator within a pipeline by replacing it with a different one or introducing new operators. This set of genetic operators allows TPOT to explore a wide range of pipeline configurations. This process is done iteratively for a specified number of generations or until a termination condition is met. After the specified number of iterations, TPOT selects the best-performing pipeline from the final population. Although, TPOT evaluates several machine learning pipelines, only the best performing pipeline is retrievable and can be extracted. This pipeline represents the optimized solution (machine learning pipeline designed) for the given dataset. The settings of TPOT used specifically in this research is discussed in Section 5.

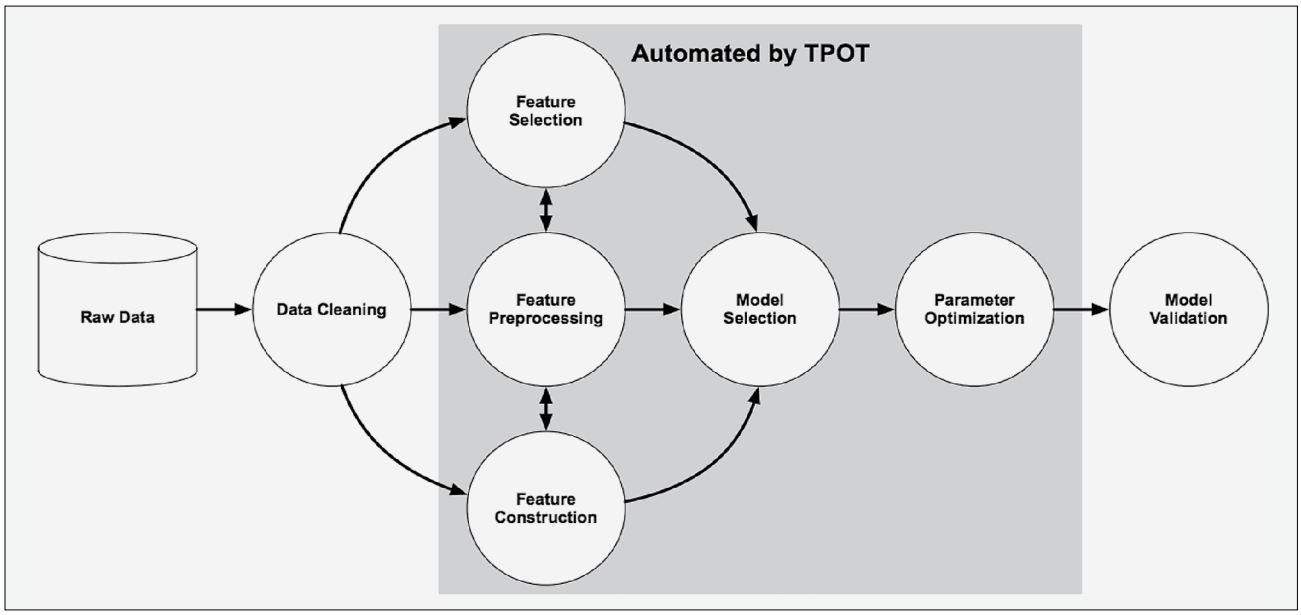

Figure 1: Workflow of TPOT

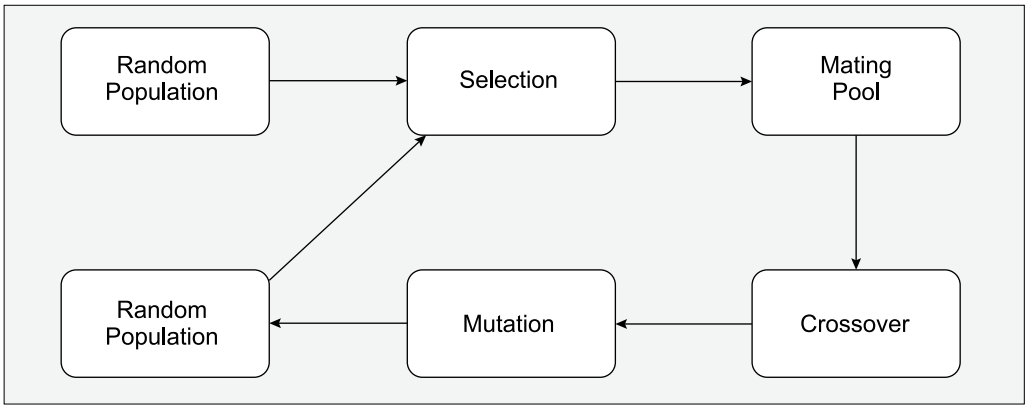

Figure 2: Workflow of genetic programming

3. Design of experiments

In the oil and gas industry, the reservoir fluids are produced to the surface as soon as the well is drilled and completed. The reservoir fluids include oil, gas and water and these must be processed at the surface. For proper planning and handling of these fluids, operators must understand the basic physics of fluid flow in the pipe or wellbore for optimal design of surface facilities (**Ganat and Hrairi, 2018**).

The TPOT AutoML library in python programming language has been used to design optimal machine learning models for various prediction tasks. Since the dataset that is to be provided as the input to the TPOT must have been framed out of some experimental results, the research done by Dvora Barnea, Ovadia Shoham, and Yehuda Taite in (**Barnea et al., 1980**) has been considered. The work presented in (**Barnea et al., 1980**) involves experiments using horizontal and inclined pipes to observe and analyze the flow patterns.

It is worth noting that the research work presented in this paper was carried out based on the datasets generated and presented by **Barnea et al. (1980)**. In all the experiments, the exact setup was replicated or they were kept similar. This study aims at depicting the way of using one of the advanced machine learning techniques, that is AutoML, for automated design of machine learning models for identifying the flow patterns in two phase (gas - liquid) systems with the help of this well-structured dataset

Barnea and other authors considered various parameters such as liquid and gas flow rates, pipe diameter, and system pressure to study their effects on flow pattern transitions. High-speed imaging techniques were utilized to capture the flow patterns accurately.

The experimental observations reveal that the researchers identified and classified different flow patterns, such as Dispersed Bubble Flow, Stratified Smooth Flow, Stratified Flow, Wavy Flow, Annular Flow, Intermittent Flow, and Bubble Flow. They analysed the characteris-

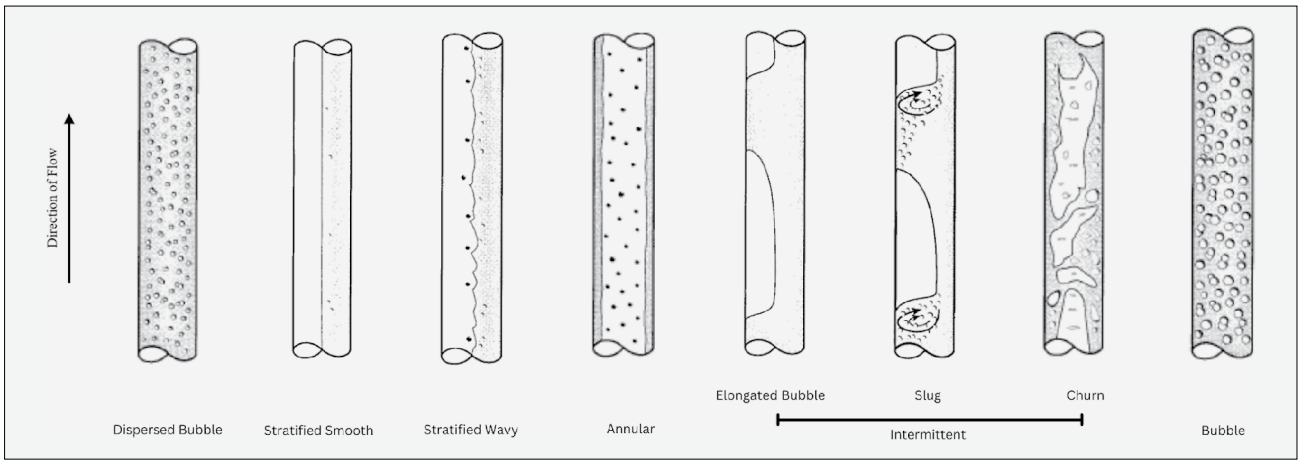

Figure 3: Flow Patterns observed in the prediction of two-phase flows (**Shoham, 2006; Khaledi et al., 2014; Hanazadeh et al., 2017; Thome, 2016; Haiyan et al., 2019; Su et al., 2022**)

tics (**Cheng et al., 2008**) and behaviour of each flow pattern (**Wu et al., 2017; Almutairi et al., 2020**) and identified the key factors that influence the transition between them (**Al-Sarkhi et al., 2016**). The study found that the transition between flow patterns is primarily influenced by the liquid and gas flow rates, as well as the pipe diameter. Additionally, system pressure was observed to have an impact on the transition process.

Based on the data collected from the existing literature, machine learning models are created to predict the flow pattern from the 2.5 cm dataset and 5.1 cm dataset. The designed models are named as FPTL_TPOT_2.5 and FPTL_TPOT_5.1, respectively, for 2.5 cm and 5.1 cm datasets, in this research. The following subsection provides a brief about the dataset description along with the explanation of the parameters, and the design methodology that involves data preparation and TPOT implementation.

4. Dataset description and pre-processing of data

The data obtained from **Barnea et al. (1980)** has several parameters which are useful to determine the flow pattern of the systems. The commonly known flow patterns are Dispersed Bubble Flow, Stratified Smooth Flow, Stratified Wavy Flow, Annular Flow, Intermittent Flow and Bubble Flow, as observed in several investigations under various settings and shown in **Figure 3** (**Shoham, 2006; Khaledi et al., 2014; Hanazadeh et al., 2017; Thome, 2016; Haiyan et al., 2019; Su et al., 2022**). This study also observed same flow patterns using datasets from **Barnea et al. (1980)**. A short description of these flow patterns can be seen in **Thome (2016)**.

Understanding two-phase flow patterns is vital in oil-gas industries, chemical engineering industries, and nuclear reactor design laboratories, etc. Accurate flow pattern identification is key for system optimization, safety, and efficiency. The features present in the dataset, considered for this study, are density of liquid, density of gas, superficial viscosity of liquid, superficial viscosity of gas, roughness of pipe, surface tension, inner diameter of pipe, system pressure, angle of inclination, superficial liquid velocity and superficial gas velocity. Of these features, the liquid and gas density, liquid and gas superficial viscosity, roughness of pipe, ST (surface tension) and ID (inner diameter of the Pipe) were observed to be constant throughout the experiments on the datasets, making them unlikely in helping the machine learning model to identify the flow pattern. The parameters that were observed to vary are System Pressure, Angle of Inclination, Superficial Liquid Velocity, and Superficial Gas Velocity. They were considered as the feature set. A short note on these parameters is presented below.

1. *System Pressure* (in N/m) - The pressure at which the two-phase flow system operates is called as the System Pressure. It affects the density, and compressibility of the fluid phases, which in turn influences the flow patterns.
2. *Angle of Inclination* (in degrees) - In the petroleum industry the oil wells can be horizontal, vertical, or slanted. As seen earlier, orientation (angle of inclination) of the wellbore or pipeline can affect the flow pattern of the multiphase fluid in the pipe and this has been considered during the machine learning model preparation.
3. *Superficial Liquid Velocity* (in m/s) - **Superficial Liquid Velocity** (V_{sl}) refers to the hypothetical flow velocity of the liquid phase in gas-liquid two-phase flow system, representing the average velocity of the liquid flowing through the system.

$$V_{sl} = \frac{Q_l}{A} \quad (1)$$

*Description of Formula in **Equation 1***: As observed in **Barnea et al. (1980)**, V_{sl} is calculated by dividing the volumetric flow rate () of the liquid by the cross-sectional area (A) of the pipe or channel carrying the flow, as shown in Equation 1 and is measured in m/s.

4. *Superficial Gas Velocity* (in m/s) **- Superficial Gas Velocity (**V_{sg}) denotes the velocity of the gas phase in gas-liquid two-phase flow system, representing the average velocity flowing through the system.

$$V_{sg} = \frac{Q_g}{A} \quad (2)$$

Description of Formula in ***Equation 2***: As observed in **Barnea et al. (1980)**, V_{sg} is calculated by dividing the volumetric flow rate (of the gas by the cross-sectional area (A) of the pipe or channel, as shown in **Equation 2** and is measured in m/s.

With the considered feature set, the flow patterns (representing various classes) are to be determined, making it a classification problem. Additionally, the work of **Barnea et al. (1980)**, consisted of experiments on a 10 m pipe length under two different diameters such as 2.5 cm and 5.1 cm. These measurements indicate the cross-sectional dimensions of the pipes through which the two-phase flow was conducted. The selection of different pipe diameters was considered as an important aspect of the experiment as it allows the investigation of the influence of pipe size on flow pattern transitions. The number of data points present in the 2.5 cm dataset was 2695 and the 5.1 cm dataset consisted of 2983 data points. This paper uses the TPOT library in designing machine learning model pipelines based on the datasets obtained for two different pipe diameters 2.5 cm and 5.1 cm, to solve the classification problem of identifying different flow patterns, with the considered feature set.

4.1. Data preparation

Data preparation is a crucial step in machine learning model creation that involves transforming raw data into a format that is suitable for training a model. It includes a series of techniques and operations aimed at organizing the data to enhance the performance and accuracy of machine learning algorithms. The 2.5 cm and 5.1 cm datasets were observed to be complete and devoid of any missing or incomplete data points. Handling of imbalanced data, Standardization and One Hot Encoding (**Cerda et al., 2018**) were performed in order to ensure that the machine learning model is able to learn the data accurately.

4.1.1. Handling of imbalanced data

In imbalanced data sets, the classes are unequally represented. In these datasets, one or more classes would have fewer or higher number of instances than the other classes. This is a common issue in many real-world datasets and also with the dataset of our interest. Usage of an Imbalanced dataset leads the machine learning models to be biased towards the majority class, as they prioritize accuracy. Consequently, the model may achieve high accuracy for the majority, while performing poor for the minority class.

To handle this, the resampling techniques such as Under-sampling and Over-sampling were used. Under-sampling refers to bringing the number of instances of the majority class to a number of minority levels and Over-sampling refers to the creation of synthetic samples of the minority class to match the number of majority class instances. The former might reduce the number of samples present in the dataset, thereby compromising the robustness of the model. The latter however is beneficial as it adds extra samples.

For the datasets used in the experiments of this paper, the over-sampling of data was carried out using the Synthetic Minority Over-Sampling Technique (SMOTE) library. The SMOTE generates synthetic samples by interpolating the neighbouring instances of the minority class. This was done by selecting a random instance from the minority class, identifying its k nearest neighbours, and creating new samples between the selected instance and its neighbours by interpolation. **Table 1**, represents the number of classes present in 2.5 cm and 5.1 cm datasets before and after using SMOTE. The symbol "*" in **Table 1** denotes that there was no presence of Bubble Flow in the 2.5 cm dataset. By introducing synthetic samples, SMOTE effectively increases the number of instances in the minority class, providing the model with a more balanced training dataset. This helped in overcoming the bias towards the majority class and allows the machine learning algorithm to learn from a more representative set of samples/instances.

Table 1: Result of using SMOTE for handling Imbalanced dataset

Sno	Flow Pattern	For 2.5 cm dataset		For 5.1 cm dataset	
		Original count	Resample Count	Original count	Resample Count
1	Dispersed Bubble Flow	270	1384	325	1521
2	Stratified Smooth	64	1384	76	1521
3	Stratified Wavy Flow	413	1384	465	1521
4	Annular Flow	563	1384	470	1521
5	Intermittent Flow	1384	1384	1523	1521
6	Bubble Flow	0*	0*	470	1521

4.1.2. Standardization

Standardization plays a key role in bringing the features present in the feature set to a similar scale, preventing bias, and improving the efficiency of the learning process. It transforms the data in a way that each feature has zero mean and unit variance. As shown in the **Equation 3**, standardization is achieved by subtracting the

Table 2: Output of TPOT for 2.5 cm and 5.1 cm Datasets

For 2.5 cm dataset		For 5.1 cm dataset	
Generation number	**Best CV Score**	**Generation number**	**Best CV Score**
1	0.9783338546554061	1	0.9726027397260275
2	0.9783338546554061	2	0.9726027397260275
3	0.9783338546554061	3	0.9726027397260275
4	0.9783338546554061	4	0.9791780821917808
5	0.9841257428839537	5	0.9791780821917808
6	0.9841257428839537	6	0.9797260273972602
7	0.9841257428839537	7	0.9797260273972602
8	0.9862840162652488	8	0.9830136986301371
9	0.9862840162652488	9	0.9846575342465753
10	0.9862840162652488	10	0.9852054794520548
Best Pipeline: MLPClassifier (PCA (FastICA (input_matrix, tot = 0.9), iterated_power = 8, svd_solver = randomized), alpha = 0001, learning_rate_init = 0.01)		**Best Pipeline:** GradientBoostingClassifier (FastICA (input_matrix, tot = 1.0), learning_rate = 0.5, max_depth = 3, max_features = 0.45, min_samples_leaf = 9, min_samples_split = 16, n_estimator = 100, subsample = 0.6500000000000001)	

mean of each feature from the data and dividing it by the standard deviation (**Ali et al., 2014**). By standardizing the data, all the features are transformed to a similar scale, which was particularly useful for algorithms that were sensitive to the magnitude of the input variables.

$$x_{\text{standardized}} = \frac{x - x_{\text{mean}}}{x_{\text{standard deviation}}} \quad (3)$$

Standardization technique is a vital step in machine learning as it cleans and transforms raw data into a suitable format for training models. It improves the quality and reliability of the data, ensures compatibility with algorithms, and enhances the overall performance of machine learning models. This was carried out for both 2.5 cm and 5.1 cm datasets.

4.1.3. One Hot Encoding

One-Hot Encoding is a crucial data pre-processing technique in the field of machine learning and is particularly relevant when dealing with categorical data or features that don't have a natural ordinal relationship (**Cerda et al., 2018**). It works by creating a binary matrix representation of data, where each unique category is represented by a distinct binary column (or "bit"). The term "one-hot" stems from the fact that only one bit is "hot" (set to 1) for a given category, while all others are "cold" (set to 0).

In the dataset obtained, the feature, "Angle of Inclination" was observed to have angles in degrees. It is essential to apply one hot encoding for this feature since it preserves the information and ensures that each unique angle is treated as a distinct category by the machine learning algorithm. If not One-Hot Encoded, the algorithm might interpret the angles as continuous numerical values, leading to misinterpretation and inaccurate results. One-Hot Encoding helps in avoiding such misinterpretations.

These steps (Handling of Imbalanced data, Standardization and One hot encoding) are performed in order and the prepared data was passed to the TPOT for identification of the right machine learning model.

5. Experimental result and discussion

This study, leverages a contemporary machine learning tool, called AutoML, to identify the most appropriate machine learning model for predicting flow patterns using the datasets obtained from the experiments conducted by **Barnea et al. (1980)** for two distinct pipe diameters of 2.5 cm and 5.1 cm. Both of these datasets were cleaned initially to remove the redundant data present and then were put to data pre-processing. All the features in the feature set were standardized and scaled to prevent bias towards any feature based on its magnitude.

Any of the flow patterns, Dispersed Bubble, Stratified Smooth, Stratified Wavy, Annular, Intermittent, and Bubble were to be predicted out of the features available at the 2.5 cm and 5.1 cm datasets using the models designed by the TPOT. TPOT addresses this as a classification problem to classify the data points to its flow pattern (class) effectively. The results obtained for the 2.5 cm and 5.1 cm datasets are presented in **Table 2**, where the TPOT was configured to run for 10 generations (termination condition) with a population size of 20 and with 5-fold Cross Validation (CV). The CV designed to be 5 denotes that the model is trained and tested five times, each time with a different subset serving as the test set. The CV score provides an overall assessment of the model's consistency and generalization ability.

An in-depth explanation for the values of parameters, obtained from the TPOT library are as follows. From

Table 3: The accuracy measurements of the models

Model Name	Accuracy Obtained (%)		
	For 2.5 cm dataset	For 5.1 cm dataset	Average
Logistic Regression	57.35	50.09	53.72
K Nearest Neighbors	56.83	54.2	55.56
Nave Bayes Classifier	57.94	65.33	61.64
Decision Tree	64.22	59.35	61.79
XGB Classifier	62.25	55.36	58.81
Neural Network classifier	69.04	51.67	60.36
SVC	59.17	56.00	57.59
TPOT designed Models	FPTL_TPOT_2.5: 97.66	FPTL_TPOT_5.1: 98.09	97.88

Table 2, it can be viewed that the resulting pipeline for the 2.5 cm dataset is a Multi-Layer Perceptron (MLP) classifier, a powerful neural network architecture often employed for classification. In the pre-processing phase, which was separately carried out by the TPOT, the dimension of data was reduced through Principal Component Analysis (PCA) and Fast Independent Component Analysis (FastICA). The PCA component was observed to be configured in order to retain 90% of the variance in the data, and specific parameters were meticulously tailored to the dimensionality reduction techniques. Furthermore, the MLP classifier was fine-tuned with a regularization strength (alpha) set at 0.0001 and an initial learning rate of 0.01.

On the other hand, the 5.1 cm dataset was best classified using the Gradient Boosting Classifier, a well-established ensemble learning algorithm utilized for classification tasks. In the pre-processing phase, autonomously managed by TPOT, feature extraction and reduction techniques were applied. Independent Component Analysis (ICA) was used to reduce the dimensionality of the data, and in this instance, it was observed to achieve a total retention of 1.0. Moreover, the selected Gradient Boosting Classifier was fine-tuned with specific hyperparameters. The learning rate was set to 0.5, which governs the step size during optimization. The maximum depth of the individual decision trees in the ensemble was limited to 3. Additionally, a maximum of 45% of features considered for each split decision, denoting that algorithm would randomly select this subset of features, which could help improve the generalization of the model by making the trees less deep. The value of 9 was kept as samples required in a leaf node for a minimum. It means that if a node has fewer than 9 data points after a split, the tree would not continue to split it, and it would become a leaf node. The minimum samples necessary to split an internal node was set at 16, meaning that an internal node must have at least 16 data points to be considered for further splitting. If a node has fewer than 16 data points, it would not be split, and the decision tree-building process would proceed to other nodes that met this criterion. The ensemble comprises 100 boosting stages. A subsample of approximately 65% of the dataset was used for training each tree in the ensemble.

Accuracy is a performance metric used in classification tasks to gauge how effectively any machine learning model correctly assigns data points to their respective categories or classes. It is computed as the ratio of the number of data points that the model correctly classifies to the total number of data points in the dataset. In practical terms, a higher accuracy score signifies that the model is more proficient at making correct classifications, indicating its ability to accurately predict the class labels of data points. With the machine learning models designed by the TPOT, the accuracy scores of 97.66% and 98.09% were observed for the 2.5 cm and 5.1 cm datasets respectively. These higher accuracy scores were achieved using TPOT by carefully designing the machine learning models which were best suited for the datasets considered along with their best hyperparameter settings.

A comparative study of the FPTL_TPOT models with other existing standard and similar machine learning models were carried out in two phases, as listed below.

Phase I: To compare with standard machine learning models.

Phase II: To compare with similar two-phase flow prediction models.

5.1 Comparative study with standard machine learning models

To compare the performance of the FPTL_TPOT models, several standard machine learning algorithms were considered as base models and were directly applied to test the accuracy of classifying the flow patterns. The standard machine learning models considered were Logistic Regression (**Wright, 1995**), K Nearest Neighbors (**Cover et al., 1967**), Naïve Bayes Classifier (**Rish, 2001**), Decision Tree (**Magee, 1964**), XGB classifier (**Agarwal et al., 1994**), Neural Network (**Bishop, 1994**) and Support Vector Classifier (SVC) (**Tong et al., 2001**). These supervised machine algorithms follow different mathematical methods to understand the features provided in the datasets and to classify the input based on this feature set. These models had been identified as effective but different in classification and regression tasks

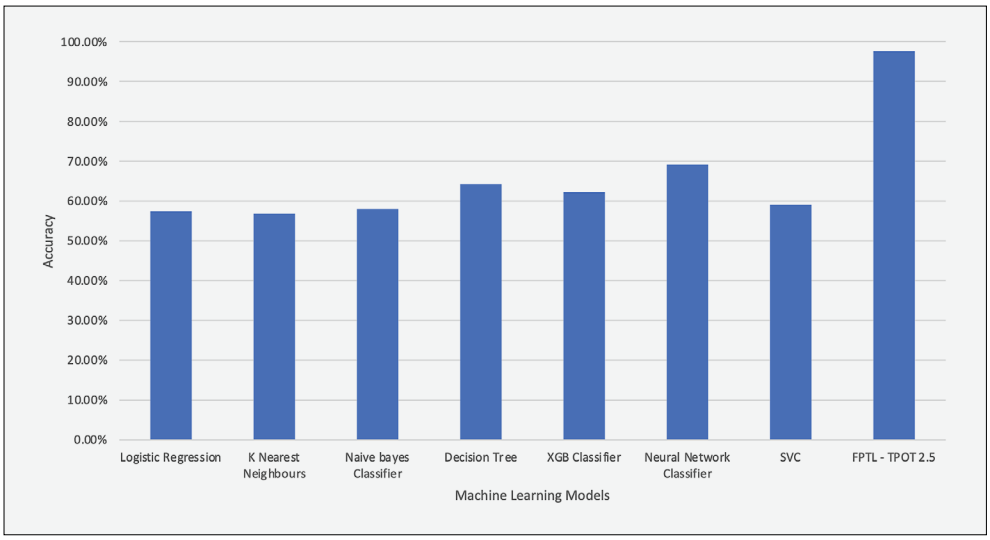

Figure 4: Comparative Results of machine learning models for 2.5 cm dataset

(**Gajula et al., 2024; Sawe et al., 2024; Nakhipova et al., 2024, Zhou et al., 2024; Hoque et al., 2024; Oyewole et al., 2024; Amaya-Tejera et al., 2024**). Although, the working of all these models and the TPOT AutoML model differs, any of them can be employed in identifying the flow patterns, as they are effective in classification and regression tasks. The performance of the models can be determined by measuring how accurately they predict the right flow pattern.

The experiments were carefully designed to implement all these seven models (Logistic Regression, K Nearest Neighbors, Nave Bayes Classifier, Decision Tree, XGB Classifier, Neural Network classifier and SVC) along with the two newly designed TPOT models; (FPTL_TPOT_2.5 and FPTL_TPOT_5.1). **Table 3** presents the accuracy comparison of the standard machine learning models along with the TPOT designed models. These results are visualized in **Figure 4** and **Figure 5**.

The results for the classification of 2.5 cm dataset (given in **Table 3** and **Figure 4**) show that, among all the models tested, the FPTL_TPOT_2.5 model outperformed the rest by a significant margin, achieving an impressive accuracy score of 97.66%. This remarkable result indicates that FPTL_TPOT_2.5 is exceptionally well-suited for handling the specific characteristics of the 2.5 cm dataset, making it the clear frontrunner in terms of accuracy.

On a one-to-one comparison of the FPTL_TPOT_2.5 model with the standard machine learning models, the following inferences are made.

1) The Logistic Regression model, which has produced an accuracy of 97% in the task of detecting brain tumour from MRI images (**Gajula et al., 2024**), achieved an accuracy of 57.35% only in the task of flow pattern prediction from the 2.5 cm dataset. However, the FPTL_TPOT_2.5 model has shown 40.31% of improvement in the accuracy of flow pattern prediction comparing to the Logistic Regression model.
2) On a phishing web page detection task, the K-Nearest Neighbor model has demonstrated better than other models with an accuracy of 97% (**Sawe et al., 2024**). However, this model could achieve only 56.83% of accuracy in the flow pattern prediction task discussed here. The FPTL_TPOT_2.5 model showed 40.83% improvement in the accuracy of flow pattern prediction, comparing to K-Nearest Neighbor model.
3) Though recently, the Nave Bayes Classifier proved its capability in predicting student achievements by assessing their educational performances (**Nakhipova et al., 2024**), its accuracy in predicting the flow pattern for the 2.5 cm dataset is 57.94%. The FPTL_TPOL_2.5 outperformed Nave Bayes Classifier with the performance improvement of 39.72%.
4) The research work of (**Zhou et al., 2024**) compared the performance of Decision Tree model with XGB classifier and Random Forest, on the task of estimating geo-polymer concrete compressive strength. In their study, the Decision Tree model is used as base learner and the other two models as super learners. The authors showed that the Decision Tree model fails in outperforming other two models. The similar trend was shown by Decision Tree model in flow pattern prediction task also, where the FPTL_TPOT_2.5 model performed better than the Decision Tree model with a 33.44% improvement in the accuracy.
5) The XGB classifier, which has demonstrated an accuracy of 94.74% in predicting breast cancer (**Hoque et al., 2024**), was outperformed by the FPTL_TPOT_2.5 model in the flow pattern prediction task for the 2.5 cm dataset. The

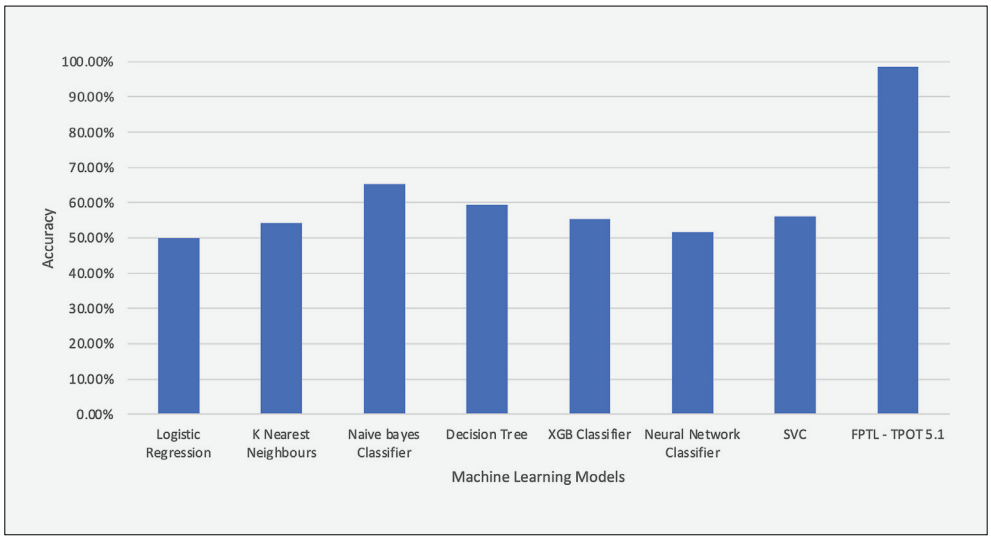

Figure 5: Comparative results of machine learning models for 5.1 cm dataset

performance difference between the models is 35.41%.

6) The Neural Network Classifier models showed accuracies of 90%, 93% and 95%, etc. in various prediction tasks in the stock market scenario. This fact is available in (**Oyewole et al., 2024**). However, the Neural Network Classifier could demonstrate an accuracy of 69.04% only for the flow pattern prediction with the 2.5 cm dataset. This has a performance difference of 28.62% comparing to the accuracy of the FPTL_TPOT_2.5 model. It is worth noting that among all the seven standard machine learning models, the Neural Network Classifier is the top performer.
7) A Support Vector Classifier added with an innovative approach of randomly selecting the training samples from the dataset was presented in the work of **Amaya-Tejera et al., (2024)**. The SVC presented showed a classification accuracy of 89% and 94.9%, for a multiclass classification, with different datasets. However, in the case of flow pattern prediction the SVC could demonstrate an accuracy of 59.17%, which is 38.49% less than that of the FPTL_TPOT_2.5 model.

Similarly, for the 5.1 cm dataset, the results (given in **Table 3** and **Figure 5**) prove accuracy that the FPTL_TPOT_5.1 model remains the top performer, achieving a higher accuracy of 98.09% on the 5.1 cm dataset. The Naive Bayes Classifier also notably achieved an accuracy of 65.33%, making it one of the top traditional models for this dataset. Decision Tree shows an accuracy of 59.35% and other models display modest scores but still trailed behind the FPTL_TPOT_5.1 model. It was observed from the one-on-one comparison that the FPTL_TPOT_5.1 model showed 48.00%, 43.81%, 32.76%, 38.74%, 42.73%, 46.42% and 42.09% of performance improvement compared to the Logistic Regression, K Nearest Neighbors, Nave Bayes Classifier, Decision Tree, XGB Classifier, Neural Network classifier and SVC models, respectively.

These results emphasize the critical importance of selecting the right machine learning algorithm when working with specific datasets, as it can significantly impact the accuracy and effectiveness of the model.

5.2 Comparative study with two-phase flow prediction models

This section presents a comparison of the FPTL_TPOT models with a few state-of-the-art similar models used for gas-liquid two-phase flow pattern prediction. The summary of the inferences is presented in **Table 4**, and are explained in detail below.

Mask et al. (2019) have proposed a set of machine learning based models for predicting gas-liquid flow pattern. The machine learning models used in this work are Random Forest (RF), AdaBoost and XGBoost with reported flow pattern prediction accuracy of 92.3%, 92% and 93.7%, respectively. These accuracies were compared to the TPOT designed models (FPTL_TPOT_2.5 and FPTL_TPOT_4.1) of 97.66% and 98.09%. The performance improvements achieved by the FPTL_TPOT_2.5 model was 5.33%, 5.66% and 3.96, respectively, while that of FPTL_TPOT_5.1 model, respectively was 5.76%, 6.09% and 4.39%.

Huang (2024), reported the performance of three machine learning models (support vector machine (SVM), K-Nearest Neighbor (KNN) and Random Forest (RF)) in the task of predicting gas-liquid flow patterns. This study discussed the performances of these machine learning models in two stages. In Stage 1, the classical SVM, KNN and RF machine learning models were used for the flow pattern prediction. The accuracy attained by these models were 73%, 91.9% and 94.6%, respectively.

Table 4: Comparison FPTL_TPOT models with state-of-the-art models

Reference Paper	Machine Learning Model	Accuracy (%)	Performance Improvement in (%)	
			By FPTL_TPOT_2.5 (Accuracy: 97.66%)	By FPTL_TPOT_5.1 (Accuracy: 98.09%)
(Mask et al., 2019)	RF	92.30	5.33	5.76
	AdaBoost	92.00	5.66	6.19
	XGBoost	93.70	3.96	4.39
(Huang et al., 2024)	SVM	73.00	24.66	25.09
	KNN	91.90	5.76	6.19
	RF	94.60	3.06	3.49
	Improved KNN	99.50	-1.84	-1.41
	Improved RF	97.50	0.16	0.59
(Hernandez et al., 2019)	Decision Tree for intermittent flow	86.32	11.34	11.77
	Decision Tree for annular flow	49.11	48.55	48.98
(**Ezzatabadipour et al., 2017**)	Deep Learning model	85.97	11.69	12.12
(**Loyola-Fuentes et al., 2022**) (for ethanol)	K-nearest neighbors	82.80	14.86	15.29
	Multilayer perceptron	83.90	13.76	14.19
	Random forest	82.20	15.46	15.89
(**Arteaga-Arteaga et al., 2021**)	Extra Tree	97.00	0.66	1.09
(**Guillen-Rondon et al., 2018**)	SVM	97.00	0.66	1.09

It was evident that the TPOT models are performing better that these models with improved accuracies. The accuracy improvements are (24.66%, 5.76%, 3.06%) and (25.09%, 6.19%, 3.49%) by the FPTL_TPOT_2.5 and FPTL_TPOT_5.1 models, respectively. In Stage 2, improved KNN and RF models were designed where the input features went through a correlation analysis to identify more influencing features from the dataset. A linear interpolation procedure was applied on the dataset, as preprocessing. The accuracy reported by the improved KNN and RN models are 99.5% and 97.5%. The accuracy differences between these models and the FPTL_TPOT_2.5 and FPTL_TPOT_5.1 models are (-1.84%, 0.16%) and (-1.41%, 0.59%), respectively. Though, the performance differences are marginal, this comparative study adds an insight to the future works on this study to add a feature reduction technique as part of data pre-processing stage to improve the prediction accuracies of the FPOT_TPOT models.

Hernandez et al. (2019) built a decision tree based model for flow pattern prediction under different two-phase flow conditions. This tree based model secured 86.32% and 49.11% accuracies for the intermittent and annular flow patterns, respectively. These accuracies are less than the accuracies achieved by the FPTL_TPOT models. The performance differences are presented in **Table 4.**

Ezzatabadipour et al. (2017) proposed a deep learning based algorithm to predict flow patterns in two phase flow under different fluid properties and pipe conditions. The initial model showed 83.87% of prediction accuracy, and an improved model had shown an accuracy of 85.97%. Comparing to the improved model the FPTL_TPOT_2.5 and FPTL_TPOT_5.1 models showed performance differences of 11.69% and 12.12%, respectively.

Loyola-Fuentes et al. (2022), in their study, trained three classification algorithms viz K-nearest neighbors, multilayer perceptron and random forest for flow pattern classification and generated a flow pattern map with boundaries between the slug/plug and annual flows. The accuracy scores reported for these algorithms are 82.8%, 83.9%, 82.2% and 75.8%, 77.1%, 75.6% under different parameter settings for ethanol and FC-72, respectively (**Loyola-Fuentes et al., 2022**). It can be observed that the FPTL_TPOT models are far better than these algorithms with higher accuracy of predictions, with good performance difference. The results are presented in **Table 4**.

A comparative study on the performance of nine different machine learning models on predicting the gas-liquid two-phase flow patterns in pipes (**Arteaga-Arteaga et al., 2021**), identified the Extra Tree (ET) model as the best performing model with a prediction accuracy of 97.00%, which is less than but closer to the accuracies 97.66% and 98.09% of the FPTL_TPOT models. In (**Guillen-Rondon et al., 2018**), an optimized Support Vector Machine (SVM) classifier was designed for gas-liquid two-phase flow pattern prediction. With three flow patterns (dispersed, segregated and intermittent) prediction system, SVM achieved a 97.00% accuracy, similar to the ET model of **Arteaga-Arteaga, et al. (2021)**, which is less than but closer to the FPTL_TPOT models. The summary of above inferences is presented in **Table 4**.

Table 5: Test cases and prediction for 2.5 cm dataset

Test Case	Pressure	Ang (In degrees)	Vsl	Vsg	FPTL_TPOT_ 2.5 Prediction
1	101400	0	0.04	1.5	Stratified Smooth
2	102402	20	0.0	16.8	Stratified Wavy
3	101475	0	0.25	10	Annular
4	103200	-50	0.1	9.5	Annular
5	103830	90	0.67	9.67	Intermittent
6	101367	0	0.004	0.04	Stratified Smooth
7	102748	-1	0.06	0.4	Stratified Wavy
8	102395	20	2.99	0.095	Dispersed Bubble
9	102843	-5	1.6	0.04	Dispersed Bubble
10	104020	-80	1.68	26.16	Intermittent

Table 6: Test cases and prediction for 5.1 cm dataset

Test Case	Pressure	Ang (In degrees)	Vsl	Vsg	FPTL_TPOT_ 5.1 Prediction
1	103390	-70	0.10	10.2	Stratified Wavy
2	101375	0	0.002	0.025	Stratified Smooth
3	103598	50	0.002	14.30	Annular
4	103080	-50	2.637	0.02	Dispersed Bubble
5	103853	70	1.517	0.092	Bubble
6	102370	-5	0.001	0.016	Stratified Smooth
7	103836	70	0.0	0.15	Intermittent
8	101999	5	0.025	16	Annular
9	104308	90	0.99	0.09	Bubble
10	102080	10	0.002	16	Stratified Wavy

In order to further validate the FPTL_TPOT_2.5 and FPTL_TPOT_5.1 models with sample test cases, few sample inputs are given to the models and the corresponding output predicted by them are presented in **Table 5** and **Table 6**.

6. Conclusions

The prediction of flow patterns is more crucial in the oil and gas industries, as they are important for their operational efficiency and system integrity. This study investigates the automated design of machine learning models for high accurate flow pattern prediction for two-phase flows. Leveraging data from Barnea's research, the TPOT tool (an AutoML tool utilizing genetic programming) was employed in designing optimal machine learning models for flow pattern prediction using the datasets with 2.5 cm and 5.1 cm diameters of pipe. The models designed by the TPOT tool for the 2.5 cm and 5.1 cm datasets were named as FPTL_TPOT_2.5 and FPTL_TPOT_5.1, respectively. A comparative study of these models with other standard machine learning models revealed that the FPTL_TPOT_2.5 and FPTL_TPOT_5.1 models are performing superior in predicting the flow pattern, by achieving remarkable prediction accuracies of 97.66% and 98.09%, respectively, for the given datasets. The performance of the FPTL_TPOT models were also compared with few state-of-the-art machine learning models designed for gas-liquid two-phase flow pattern prediction. Except for a case with improved KNN, the FPTL_TPOT models demonstrated improved prediction accuracies for other state-of-the-art models. The performance differences observed were in the range of 0.59% to 25.09%.

These findings hold significant implications for industrialists and researchers, facilitating the precise prediction of flow patterns crucial for operational design and analysis in the petroleum industry. However, it's important to note that the generalizability of the findings may be limited by the specific characteristics of the datasets used in this study. Despite this, the outcomes presented herein offer valuable insight and pave the way for further research works.

Future investigations could include exploring alternative machine learning models and evolutionary algorithms with diverse parameter settings to delve deeper into the prediction of multi-phase gas-liquid flow patterns, towards designing a generalized system which can

work for any given dataset. Incorporation of a suitable dimensionality reduction technique, to identify more relevant features from the dataset, towards increasing the prediction accuracy of the FPTL_TPOT models is also an interesting future scope. Integrating other Artificial Intelligence methodologies into this research can contribute meaningfully to the ever-evolving landscape of the petroleum industry, for enhancing their operational efficiency and system reliability.

7. References

Abduvayt, P., Manabe, R. and Arihara, N. (2003): Effects of pressure and pipe diameter on gas liquid two-phase flow behavior in pipelines. In the proceedings of SPE annual technical conference and exhibition, Denver, Colorado, USA, 1-15.

Agarwal, A. K., Wadhwa, S., & Chandra, S. (1994): XGBoost a scalable tree boosting system. Journal of Association of Physicians India, 42, 8, 665.

Agash, U., Ramya, G. R. and Jeyakumar, G. (2023): Effective link prediction in complex networks using differential evolution based extreme gradient boosting algorithm. In the proceedings of advanced network technologies and intelligent computing. Communications in computer and information science (CCIS). Springer. 1797, 143-154.

Ali, P. J. M., Faraj, R. H., & Koya, E. (2014): Data normalization and standardization: A technical report. *Machine Learning Technical Report,* 1, 1, 1–6.

Almutairi, A., Al-Alweet, F.M., Alghamdi, Y. A., Almisned, O. A. and Alothman, O. Y. (2020): Investigating the characteristics of two-phase flow using electrical capacitance tomography (ECT) for three pipe orientations. Processes, 8, 1, 51.

Al-Sarkhi, A., Duc, V., Sarica, C. and Pereryra, E. (2016): Upscaling modeling using dimensional analysis in gas-liquid annular and stratified flows. Journal of petroleum science and engineering. 137, 240-249.

Amaya-Tejera N, Gamarra M, Vélez JI and Zurek E (2024): A distance-based kernel for classification via Support Vector Machines. Frontiers in Artificial Intelligence, 7, 1287875.

Anusha, J., Rekha, V. S. and Sivakumar, P. B. (2015): A machine learning approach to cluster the users of stack overflow forum. In the proceedings of artificial intelligence and evolutionary algorithms in engineering systems. 325, 411–418.

Arteaga-Arteaga, HB., Mora-Rubio, A., Florez, F., Murcia-Orjuela, N., Diaz-Ortega, CE., Orozco-Arias, S., delaPava, M., Bravo-Ortíz, MA., Robinson, M., Guillen-Rondon, P. and Tabares-Soto, R. (2021): Machine learning applications to predict two-phase flow patterns. PeerJ Computer Science. 29, 7, e798.

Attia, M., Mahmoud, M. A., Abdulraheem, A., and Al-Neaim, S. A. (2013): Evaluation of the pressure drop due to multi phase flow in horizontal pipes using fuzzy logic and neural networks. SPE middle east oil and gas show and conference, SPE-164278-MS.

Barjouei, H. S., Ghorbani, H., Mohamadian, N., Wood, D. A., Davoodi, S. and Moghadasi, J. (2021). Prediction performance advantages of deep machine learning algorithms for two phase flow rates through wellhead chokes. Journal of petroleum exploration and production, 11, 1233-1261.

Barnea, D., Shoham, O., Taitel, Y. and A. E. Dukler. (1980): Flow pattern transition for gas-liquid flow in horizontal and inclined pipes. International Journal of Multiphase Flow, 6, 217-225.

Batchu, R. K. and Seetha, H. (2021): A generalized machine learning model for DDoS attacks detection using hybrid feature selection and hyperparameter tuning. Computer networks. 200.

Bishop, C. M. (1994): Neural networks and their applications. Review of scientific instruments, 65, 6, 1803-1832.

Cerda, P., Varoquaux, G. & Kégl, B. (2018): Similarity encoding for learning with dirty categorical variables. Mach Learn 107, 1477–1494.

Cheng, L., Ribatski, G. and Thome, J. R. (2008): Two-phase flow patterns and flow-pattern maps: fundamentals and applications. Applied mechanics review. 61, 5, 050802.

Cover, T., & Hart, P. (1967): Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13, 1, 21-27.

Duru, U. I., Kwesi Wayo, D. D., Ogu, R., Cyril, C. and Nnani, H. (2022): Computational analysis for optimum multiphase flowing bottom-hole pressure prediction. Transylvanian review, Computers and information science. 30, 2.

Ezzatabadipour, M., Singh, P., Robinson, M. D., Guillen-Rondon, P. and Torres, C. (2017): Deep Learning as a Tool to Predict Flow Patterns in Two-Phase Flow. arXiv:705.07117.

Gajula, S., & Rajesh, V. (2024): An MRI brain tumour detection using logistic regression-based machine learning model. International Journal of System Assurance Engineering and Management, 15, 1, 124-134.

Ganat, T. A. and Hrairi, M. (2018): A new choke correlation to predict flow rate of artificially flowing wells. Journal of petroleum science and engineering. 171, 1378-1389.

Guillen-Rondon, P., Robinson, M.D., Torres, C. and Pereya, E. (2018): Support Vector Machine Application for Multiphase Flow Pattern Prediction. arXiv:1806.05054.

Hafsa, N., Rushd, S. and Yousuf, H. (2023): Comparative performance of machine-learning and deep-learning algorithms in predicting gas-liquid flow regimes. Processes. 11, 1, 177.

Haiyan, W., Chunsheng, W., Yuxing, L. and Xiaohua, C. (2019): Flow-pattern-prediction models used for gas-liquid two-phase flow. Journal of Oil & Gas Storage and Transportation, 1, 1, 55-60.

Hanazadeh, P., Eshraghi, J., Nazari, Y., Yousefpour, K. and Behabadi, M. A. A. (2017): Light oil-gas two-phase flow pattern identification in different pipe orientations: An experimental approach. Scientia Iranica, Transactions B: Mechanical Engineering. 24, 5, 2445 – 2456.

Hernandez, J. S., Valencia, C., Ratkovich, N., Torres, C. F., and Munoz, F. (2019): Data driven methodology for model selection in flow pattern prediction, Heliyon, 5, 11, e02718.

Hoque, R., Das, S., Hoque, M., & Haque, E. (2024). Breast Cancer Classification using XGBoost. World Journal of Advanced Research and Reviews, 21, 2, 1985-1994.

Huang, Z., Duo, Y. and Xu, H. (2024): Prediction of two-phase flow patterns based on machine learning. Nuclear Engineering and Design, 421, 113107.

Jahanandish, I., Salimifard, B., and Jalalifar, H. (2011): Predicting bottomhole pressure in vertical multiphase flowing wells using artificial neural networks. Journal of petroleum science and engineering, 75, 336-342.

Jayakumar, V. and Raju, R. (2011): A multi-objective genetic algorithm approach to the probabilistic manufacturing cell formation problem. South african journal of industrial engineering, 22, 1, 199-212.

Khaledi, H. A., Smith, I. E., Unander, T. E., and Nossen, J. (2014): Investigation of two-phase flow pattern, liquid holdup and pressure drop in viscous oil–gas flow. International Journal of Multiphase Flow, 67, 37–51.

Kim, D.H., Zohdi, T.I. and Singh, R. P. (2020): Modeling, simulation and machine learning for rapid process control of multiphase flowing foods. Computer methods in applied mechanics and engineering, 371, 113286.

Le, T.T., Fu, W. and Moore, J. H. (2020): Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Bioinformatics, 36, 11, 250-256, 2020.

Lin, Z., Liu, X., Lao, L. and Liu, H. (2020): Prediction of two-phase flow patterns in upward inclined pipes via deep learning. Energy. 210, 118541.

Loyola-Fuentes, J., Pietrasanta, L., Marengo, M. and Coletti, F. (2022): Machine Learning Algorithms for Flow Pattern Classification in Pulsating Heat Pipes. Energies, 15, 1970.

Magee, J. F. (1964): Decision trees for decision making. Brighton, MA, USA: Harvard Business Review, 35-48.

Malbrel, C. A., Kale, R., Agarwal, J., & Gohari, K. (2024): Multiphase Flow Pattern and Screen Selection: Two Overlooked Parameters Essential to Reservoir Control Valve Optimization. SPE International Conference and Exhibition on Formation Damage Control, Lafayette, Louisiana, USA.

Manami, M., Seddighi, S. and Orl, R. (2023): Deep learning models for improved accuracy of a multiphase flowmeter. Measurement, 206, 112254.

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, 106370.

Muthaiah, U., Markkandeyan, S. and Seetha, Y. (2019): Classification models and hybrid feature selection method to improve crop performance. International journal of innovative technology and exploring engineering, 8, 11S2.

Nakhipova, V., Kerimbekov, Y., Umarova, Z., Suleimenova, L., Botayeva, S., Ibashova, A., & Zhumatayev, N. (2024): Use of the Naive Bayes Classifier Algorithm in Machine Learning for Student Performance Prediction. International Journal of Information and Education Technology, 14, 1, 92-98.

Olson, R. S., Bartley, N., Urbanowicz, R. J. and Moore, J. H. (2016a): Evaluation of a tree-based pipeline optimization tool for automating data science. In the proceedings of GECCO 2016. 485-492.

Olson, R. S., Urbanowicz, R. J., Andrews, P. C., Lavender, N. A., Kidd, L. C. and Moore, J. H. (2016b): Automating biomedical data science through tree-based pipeline optimization. In the proceedings of european conference in applications of evolutionary computation. 123-137.

Oyewole, A. T., Adeoye, O. B., Addy, W. A., Okoye, C. C., Ofodile, O. C., & Ugochukwu, C. E. (2024): Predicting stock market movements using neural networks: a review and application study. Computer Science & IT Research Journal, 5, 3, 651-670.

Rish, I. (2001): An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence, 3, 22, 41-46.

Rushd, S., Gazder, U., Qureshi, H. J. and Arifuzzaman, M. (2022): Advanced machine learning applications to viscous oil-water multi-phase flow. Applied sciences. 12, 10, 4871.

Sawe, L., Gikandi, J., Kamau, J., & Njuguna, D. (2024): Sentence Level Analysis Model for Phishing Detection Using KNN. Journal of Cybersecurity, 6, 2579-0072.

Shoham, O. (2006): Mechanistic modeling of gas liquid two phase flow in pipes. SPE Textbook Series, Richardson, TX.

Spandana, C., Srisurya, I.V., Nandhini, S. A., Kumar, R. P., Mohan, G. B. and Srinivasan, P. (2023): An efficient genetic algorithm-based auto ml approach for classification and regression. Intelligent data communication tech. and internet of things. 371-376.

Su, Q., Li, J. and Liu, Z. (2022): Flow Pattern Identification of Oil–Water Two-Phase Flow Based on SVM Using Ultrasonic Testing Method. Sensors. 22, 16, 6128.

Thome, J. R. (2016). The Heat Transfer Engineering Data Book III. pp-publico.

Tong, Simon & Koller, Daphne. (2001). Support Vector Machine Active Learning With Applications To Text Classification. The Journal of Machine Learning Research. 2, 45-66.

Uthayasuriyan, A., Chandran G, H., UV, K., Mahitha, S. H., & Jeyakumar, G. (2024): Performance Evaluation of Evolutionary Algorithms on Solving the Influence Maximization Problem in Social Networks. International Journal of Modern Education and Computer Science, 16(2), 83-97.

Uthayasuriyan, A., Chandran, H G., Kavvin UV., Mahitha, S.H. and Jeyakumar, G. (2023): A comparative study on genetic algorithm and reinforcement learning to solve the traveling salesman problem. research reports on computer science. Universal wise publisher, 1, 12.

Wright, R. E. (1995): Logistic regression. In L. G. Grimm & P. R. Yarnold (Eds.), Reading and understanding multivariate statistics American Psychological Association, 217–244.

Wu, B., Firouzi, M., Mitchell, T., Rufford, T. E., Leonardi, C. and Towler, B. (2017): A critical review of flow maps for gas-liquid flows in vertical pipes and annuli. Chemical engineering journal. 326, 350-377.

Yusof, K. A et al., (2024): Earthquake Prediction Model Based on Geomagnetic Field Data Using Automated Machine Learning. IEEE Geoscience and Remote Sensing Letters, 21, 1-5.

Zhou, J., Su, Z., Hosseini, S., Tian, Q., Lu, Y., Luo, H., ... & Huang, J. (2024). Decision tree models for the estimation of geo-polymer concrete compressive strength. Mathematical Biosciences and Engineering, 21, 1, 1413-1444.

SAŽETAK

Predviđanje ponašanja protoka u vodoravnim i nagnutim cijevima pomoću automatiziranoga strojnog učenja temeljenoga na metodi stabla odlučivanja

Razumijevanje dvofaznoga protoka (plin-tekućina) od velike je važnosti u naftnoj i plinskoj industriji jer izravno utječe na projektiranje opreme, kontrolu kvalitete i radnu učinkovitost. U industrijskome inženjerstvu i upravljanju procesima važno je poznavanje obrasca protoka u procesu. U ovome su istraživanju, kako bi se dobio najbolji model strojnoga učenja za navedeni skup podataka, korišten alat za optimizaciju cjevovoda temeljen na stablu odlučivanja (engl. *Tree-Based Pipeline Optimization Tool*, TPOT) i automatizirani sustav strojnoga učenja (engl. *Automated Machine Learning*, AutoML). U radu je predstavljena izrada modela predviđanja uzorka protoka pomoću TPOT-a. TPOT je primijenjen za predviđanje uzoraka protoka u cijevima promjera 2,5 cm i 5,1 cm korištenjem skupova podataka iz postojeće literature. Prije unošenja u TPOT skupovi podataka prošli su obradu neuravnoteženih podataka, standardizaciju i jednokratno kodiranje. Modeli izrađeni za skupove podataka za cijev promjera 2,5 cm i cijev promjera 5,1 cm nazvani su FPTL_TPOT_2.5 odnosno FPTL_TPOT_5.1. U radu je provedena komparativna analiza navedenih modela, ostalih standardnih modela strojnoga učenja i sličnih najnovijih modela predviđanja dvofaznoga protoka te je dan osvrt na izvedbu navedenih TPOT modela. Rezultati su pokazali da modeli izrađeni TPOT-om postižu izvanrednu točnost, s rezultatom od 97,66 % za skupove podataka za cijev 2,5 cm, odnosno 98,09 % za skupove podataka za cijev 5,1 cm. Nadalje, modeli FPTL_TPOT_2.5 i FPTL_TPOT_5.1 dali su bolje rezultate od drugih modela strojnoga učenja u smislu izvedbenih karakteristika naglašavajući učinkovitost TPOT-a u stvaranju modela strojnoga učenja za predviđanje uzorka protoka. Rezultati ovoga istraživanja znatno doprinose povećanju učinkovitosti i optimiziranju industrijskih procesa u naftnome i plinskome sektoru.

Ključne riječi:
višefazni protok, dvofazni protok, predviđanje obrasca protoka, strojno učenje, AutoML

Author's contribution

Agash Uthayasuriyan (1) (M.S. in Information Systems Student**,** Department of Multidisciplinary Graduate Engineering, Northeastern University, Boston, United States of America, Research Interest – Machine Learning, Artificial Intelligence) **f**ormulated the FPTL_TPOT_2.5 and FPTL_TPOT_5.1 for the 2.5 cm dataset and 5.1 cm datasets, performed data pre-processing and configuring of TPOT for the chosen datasets, and implemented standard machine learning algorithms (linear regression, KNN, Naive Bayes, Decision tree, XGB classifier, Neural network and SVC) to compare the performance of FPTL-TPOT.
Ugochukwu Ilozurike Duru (2) (Department of Petroleum Engineering, Federal University of Technology, Owerri, Nigeria, Research Interest - Production, Reservoir and Gas Engineering) gathered the experimental dataset compiled by Barnea, performed critical study on the features (density of liquid, density of gas, superficial viscosity of liquid, superficial viscosity of gas, roughness of pipe, surface tension, inner diameter of pipe, system pressure, angle of inclination, superficial liquid velocity, and superficial gas velocity) present in the dataset and carefully selected the constant features and varying features required for flow pattern prediction.
Angela Nwachukwu (3) (Senior Lecture, Department of Petroleum Engineering, Federal University of Technology, Owerri, Nigeria, Research interests - Reservoir Engineering, Production Engineering & Flow Assurance) formulated the idea for constructing machine learning model with higher accuracy of flow pattern prediction and provided clarity on the nature and parameters of the flow patterns (Dispersed Bubble, Stratified Smooth, Stratified, Wavy, Annular, Intermittent, and Bubble flow).
Thangavelu Shunmugasundaram (4) (Associate Professor, Department of Computer Science and Engineering, Amrita School of Computing, Coimbatore, India, Research Interest – Database technologies, Evolutionary algorithms) brought clarity on genetic programming, formulated the workflow for flow pattern prediction, gathered experimental results, and took inferences from the results for comparison.
Jeyakumar Gurusamy (5) (Professor, Department of Computer Science and Engineering, Amrita School of Computing, Coimbatore, India, Research Interest – Artificial Intelligence, Evolutionary Computing, Distributed Computing) mapped the classification problem and the flow pattern prediction problem in the Machine Learning domain, implemented the flow pattern prediction system, integrated the TPOT tool, analysed the test cases FPTL-TPOT models with samples and took the role of principal investigator of this research paper.