International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 2 Feb. 2026, pp: 338-344 www.ijaem.net ISSN: 2395-5252

Benchmark Dataset & Standardized Evaluation Suite for ML Models on Multiphase Pipeline Flow

Jeremiah Ifeanyi Okoroma, Ikechi Igwe

Date of Submission: 12-02-2026 Date of Acceptance: 25-02-2026

Abstract
The correct forecasting of multiphase flow parameters of pipeline, including the flow rates, flow regimes, is paramount to optimizing hydrocarbon production and the safety of operation of pipelines. Mechanistic models, such as OLGA are computationally heavy and need very specific inputs, and cannot be used in real-time. This paper suggests a standardized evaluation suite and benchmark dataset in machine learning (ML) models on multiphase pipeline flow that overcomes the absence of reproducible datasets and standardized evaluation procedures. The test data is a collection of synthetic simulations and field tests, which represent a wide range of pipeline configurations, fluid behavior and dynamic operating conditions. Normalization, imputation of missing values and flow regime labeling that was annotated by experts were part of preprocessing. The assessed baseline models are Random Forest (RF), Gradient Boosting (GB), Long Short-Term Memory (LSTM) and Transformer-based models. In prediction of flow rate, LSTM provided the best coefficient of determination (R^2 =0.95) but GB had the lowest root mean square error (RMSE =3.79 m^3 /day). The greatest accuracy (95.1) and F1-score (0.95) with the Transformer model were with flow regimes classification (slug, stratified, annular, and bubbly). The tests of robustness in Gaussian noise (5% of the feature standard deviation) demonstrated that LSTM had the lowest increase in RMSE (+0.37 m^3/day), then GB (+0.52 m^3/day), and then RF (+0.93 m^3/day), which demonstrates that the temporal models are more resilient to noisy inputs. These findings indicate the worth of standard datasets and evaluation protocols to the research of multiphase flow using machine learning tools through models that use the temporal dependencies and advanced architectures, which are accurate and robust enough to predict the desired results. The suggested benchmark and suite allows objective comparison of algorithms to support the reproducible and faster development of ML algorithms in pipeline operations.

Keywords: Multiphase flow, Machine learning, Flow rate prediction, Flow regime classification, LSTM, Transformer, Gradient Boosting, Benchmark dataset, Pipeline monitoring, Robustness evaluation

I. Introduction

Multiphase flow, which is one of the simultaneous flows of the gas and liquid phases in the pipeline, is one of the basic phenomena in the production of hydrocarbons and the operation of wellbores, as well as in the transport system on the surface. Proper prediction of the flow rates, pressure drops, liquid holdup and transitions between flow regimes is a key to optimizing the production efficiency, reducing the cost of operation, and ensuring the safety of the process in upstream and midstream petroleum operations (Ahmed and Islam, 2020; Zhao et al., 2021). The flow instabilities like slugging may result in extreme pressure swings, equipment wear and tear, separation inefficiencies, and unexpected shut-down. As a result, the stable modeling and observation of multiphase flow behavior is still an important issue in the contemporary pipeline engineering.

Conventional modeling methods have been based on mechanistic models based on conservation laws of mass, momentum and energy, with empirical phase interaction and frictional losses correlations. These principles are applied to commercial simulators like OLGA to simulate transient multiphase flow behavior at different operating conditions (Wang et al., 2019). Mechanistic simulators are also computationally intense, have to be calibrated extensively, and can only give physically consistent predictions, which are not always easily accessible in the field. Moreover, the accuracy of prediction could be limited by the uncertainties of the empirical correlations, particularly in the complex regimes of flow or those conditions of operation that vary rapidly.

Over the past years, machine learning (ML) has become a promising data-driven substitute of multiphase flow modeling. On the other hand, physics-based simulators, machine learning models can be trained on nonlinear correlations from previous sensor readings and gain knowledge of key flow variables quickly without having to solve the governing equations (Liu et al., 2022). Ensemble learning, deep neural networks and recurring

architectures techniques have shown high performance in the prediction of pressure gradients, liquid holdup, and flow regime classification. They are appealing in terms of online monitoring, detecting anomalies, and optimization of the production process since they can process a significant amount of real-time sensor data.

Although these virtues are present, the use of ML in multiphase flow study is limited considerably. The majority of the studies are conducted using internal research data, small-scale lab experiments, or simulation data, which are not easily repeatable or objectively comparable across different algorithms (Chen & Lee, 2020). Besides that, data splitting strategies are not always consistent, performance metrics differ, and some benchmarking frameworks are not standardized, which complicates finding out the models that are best generalized in realistic pipeline conditions. The lack of standardized assessment guidelines restricts the wider usage of ML solutions in the industrial setting where reliability and validation must be of the highest importance.

In order to close these gaps, this research suggests the benchmark multiphase flow dataset which includes various operating conditions and flow regimes, and a standardized evaluation suite which aims to provide fair and reproducible comparison of models. The assessment system involves unmistakably presented data partition techniques, regression and labeling execution measures, resiliency testing protocols, and expressive benchmark models that incorporate linear, ensemble, and profound learning. This study will contribute to the levels of reliability, transparency, and industrial usability of machine learning models in predicting and monitoring multiphase pipeline flows by creating a systematic benchmarking approach.

II. Literature Review

Machine learning (ML) has become a paradigm in the prediction and analysis of the multiphase flow behavior in pipelines. Multiphase flow systems with the simultaneous dispersal of gas, oil, and water are nonlinear and transient in nature and highly responsive to operation and geometrical parameters. Conventional mechanistic and empirical correlations are not always capable of well representing the complex flow regimes of slug, annular, stratified and bubbly flows in both changing pressure and temperature. Consequently, data-driven approaches have become more and more popular in the recent years.

Flow regimes Learning algorithms Supervised learning factors have played a significant role in classifying flow regimes and making quantitative predictions of flow parameters. Support Vector Machines (SVM), k-Nearest Neighbors (kNN), Artificial Neural Networks (ANN), and Decision Trees are examples of models that have shown great accuracy in classifications when used on experimental data (Alizadeh et al., 2021; Guo et al., 2022). The input characteristics that are generally utilized in these models include superficial gas velocity, superficial liquid velocity, inclination angle of the pipe, fluid properties and pressure gradients. ML models have been applied to predict pressure drop, liquid holdup, and rate of phase flow, and in some cases, they predict better than traditional mechanistic correlations in regression applications (Gao and Feng, 2021; Mohammad and Rafiee, 2023).

Multiphase flow modeling has been placed especially high-performing in ensemble learning techniques. Random Forest, Gradient Boosting, and XGBoost examples are some of the algorithms that optimize the use of weak learners in order to improve the generalization and minimize overfitting. According to Mohammadi and Rafiee (2023), XGBoost was more accurate in estimating pressure gradients than standalone decision trees and linear regression models. On the same note, ensemble methods have been observed to be useful when dealing with noisy sensor data and nonlinear relationships typical of field measurements (Liu et al., 2022). Their strength renders them applicable in real-time use of pipeline monitoring.

The use of deep learning has also extended the functions of ML in multiphase systems. The implementation of CNNs has been used to identify spatial characteristics of flow images and sensor arrays that can be used to classify flow regimes better (Zhang et al., 2020). In modeling time-series data of transient multiphase flows (including slugging behavior in offshore pipes), Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRU) have been found to be more successful (Wang et al., 2021). These recurrent architectures capture the temporal dependencies and variations that can be ignored by a traditional feedforward network. Even more recently, hybrid CNNLSTM models have been shown to have a higher predictive accuracy to liquid holdup and pressure fluctuation analysis (Chen and Li, 2023).

Along with the supervised methods, unsupervised learning methods have been investigated as well as pattern recognition and anomaly detection. Hidden flow patterns have been identified using clustering algorithms such as k-means and hierarchical clustering, and abnormal operating conditions have been identified without

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 2 Feb. 2026, pp: 338-344 www.ijaem.net ISSN: 2395-5252

any labels (Hosseini et al., 2022). Principal Component Analysis (PCA) and autoencoders are also used as dimensionality reduction methods to extract features of the dominant flows and enhance the efficiency of the computational processes (Liu et al., 2022).

Although these improvements are made, there are a number of restrictions in the literature. A large part of existing research is based on lab-scale experimental designs or synthetic data created with the help of Computational Fluid Dynamics (CFD) modeling and mechanistic simulators like OLGA (Alizadeh et al., 2021). Although these data sets allow isolated tests of the models, they might not be comprehensive of the real-world context, such as sensor noises, scaling phenomena and instability of multiphase flows. Therefore, synthetic data-trained models might have less generalization power in deployment in production pipelines.

Lack of standardized benchmarking protocols is also another significant issue. Various studies apply different performance measures, including accuracy, F1-score, RMSE, or R 2, without regular data divisions or cross-validation plans, so they are hard to compare directly (Guo et al., 2022). Moreover, there is limited availability of datasets because of proprietary restraints on oil and gas industry, which prevents reproducibility and joint development.

New architectures like attention-based neural networks, transformer networks are not yet studied in the context of multiphase flow. The potential benefit of transformer models that have demonstrated impressive results in sequencing capturing tasks in other directions is in modeling long-range temporal relationships in transient pipeline data (Sun et al., 2024). Nevertheless, their use in multiphase flow prediction is in its early stage with significantly little labeled datasets and high computational needs.

The interpretability of the model is an essential issue. The transparency of black-box models (especially deep neural networks) is frequently limited, and such behavior does not allow black-box models to be accepted in the pipeline operations that require safety. Recent papers have started to employ the method of explainable AI (XAI) including SHAP and LIME to gain an understanding of feature significance and pathways of decision-making in models (Rahman and Ibrahim, 2023). Such practices will be necessary to increase trust levels and make the use of ML-assisted pipeline monitoring systems acceptable to the regulator.

III. Methodology

3.1. Data Generation and Collection

In this work, the dataset is based on synthetic simulation data and real field measurements to provide the diversity, robustness, and realisticness of the multiphase pipeline flow behaviour. The integration of simulated and real-world data allows the parametric exploration to be performed in a controlled way and still maintain the complexity of the real world, including noise, disturbance and short-term effects.

- **Synthetic Data from Simulation:**

Modeled on validated multiphase flow simulators on a variety of pipeline configurations (diameters, inclinations, fluid properties) and operating conditions. Such variables measured are inlet/outlet pressures, temperature, and phase flow rates, holdup, and flow regime labels. The systematically varied superficial velocities of liquid (V_L) and gas (V_G) were changed:

(1)

Where:

- Vs = superficial velocity,
- Q = volumetric flow rate,
- A = pipe cross-sectional area.

- **Field Measurements:**

In order to improve the external validity, anonymized data related to production pipelines were included. These are actual operating conditions and they include:

- Sensor noise
- Transient disturbances (e.g., slugging events)
- Missing or irregular data intervals
- Operational variability

The field dataset is a time-series collection of pressure, temperature, and flow rate data of installed sensors on production pipelines. Integration of real-worlds data: This makes sure that models developed are not only applicable outside the controlled simulation environment but also when they are applied in practice.

3.2. Data Preprocessing

Preprocessing was done to make sure that there was integrity, consistency, and compatibility of data with machine learning algorithms. Due to the heterogeneous character of the collected dataset, several steps were taken.

3.2.1 Signal Alignment and Resampling

In converting a signal to a digital representation, there is a need to align the signal at the digital sampling point of a digital signal, and there may be a need to align the signal at other signal points.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 2 Feb. 2026, pp: 338-344 www.ijaem.net ISSN: 2395-5252

Signals of various sensors were put at the same time to synchronize with uniform time stamps. To achieve a consistent temporal resolution of all the variables, interpolation techniques were used to resample. This is to avoid temporal misalignment which may distort dynamic feature relationship.

3.2.2 Feature Normalization

In order to enhance convergence when training a model and prevent dominance of large scale variables, features were normalized using training-set statistics

(2)

where μ and σ are the mean and standard deviation of the training set.

- Imputation of missing values using domain-aware interpolation.
- Flow regime labeling curated via expert annotation and cross-validation with physical indicators.

3.2.3 Missing Value Imputation

The domain-aware interpolation methods were used to deal with missing values. Linear or spline interpolation was employed to bridge short gaps, and physically informed constraints to preserve flow continuity were applied for large gaps. This eliminates distortion of dynamic flow behavior.

3.3. Standardized Evaluation Suite

A strong evaluation suite has been created in order to make comparisons between models fairly.

3.3.1. Evaluation Protocols

- Train/Validation/Test Split: Temporal split (60% train, 20% validation, 20% test) to prevent leakage; stratified by flow regimes.
- Cross-Validation: K-fold on training data for hyperparameter search.

3.3.2. Performance Metrics

1. **Regression (flow rate prediction):**

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\bar{y} - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

(3)

Where:

- = total number of observations
- = actual value of the i-th observation
- = predicted value of the i-th observation
- = mean of the actual values ()
- = residual sum of squares (unexplained variance)
- = total sum of squares (total variance)

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\bar{y} - y_i|$$

(4)

Where:

- = total number of observations
- = actual (true) value of the i-th observation
- = predicted value of the i-th observation
- = absolute error for each observation

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\bar{y} - y_i)^2}$$

(5)

Where:

- = total number of observations
- = actual value of the i-th observation
- = predicted value of the i-th observation
- = squared error for each observation

2. Classification (flow regime prediction): Accuracy, F1-score, confusion matrices, ROC, and AUC.
3. Robustness Metrics: Sensitivity to noise and out-of-distribution generalization.

3.3.3. Baseline Models

In order to have fair benchmarking, representative machine learning algorithms at varying complexity were chosen as baseline models. Linear models were also added such as Linear Regression, Ridge Regression, and Lasso Regression to offer easy to understand reference points, as well as estimate linear relationships between input variables and flow parameters. Ensemble strategies based on trees, including Random Forest and Gradient Boosting (e.g., XGBoost), were also added because they have a high ability to deal with nonlinear interactions and high-dimensional data characteristics of multiphase flow systems.

Furthermore, neural network architecture was applied to model complicated and variable behavior. Steady-state prediction tasks were done using Feedforward Artificial Neural Networks (ANNs), and temporal dependencies of transient flow data were done using Long Short-Term Memory (LSTM) networks. State of the art attention-based systems, such as Transformer models, were also included to test their capability of learning long term temporal relations. The rationale of such a variety of baseline models is to provide a comprehensive comparative analysis of the conventional statistical approaches, ensemble learning, and the latest deep learning approaches.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 2 Feb. 2026, pp: 338-344 www.ijaem.net ISSN: 2395-5252

IV. Results

4.1. Flow rate Prediction Regression Performance Regression Performance: Prediction of flow rate.

Three typical ML models that we tested included the Random Forest (RF), Gradient Boosting (GB), and LSTM to predict the liquid flow rate (m^3 / day) based on pipeline data. The test set was 20 percent of the entire data, and the unseen operating conditions. Table 1 is a summarization of the performance.

Table 1: Flow rate Prediction Regression Metrics.

Model	MAE (m^3/day)	RMSE (m^3/day)	R^2
Random Forest	3.25	4.12	0.92
Gradient Boosting	2.87	3.79	0.94
LSTM	2.95	3.65	0.95

The regression findings indicate that the three models are all capable of predicting liquid flow rates and LSTM (0.95) has the highest R^2 (excellent fit to the data). The lowest RMSE (3.79 m^3/day) was documented in Gradient Boosting as it indicates a high overall accuracy in a variety of operating conditions. Although slightly LSTM was able to capture transient variations compared with Gradient Boosting, under rapidly changing flow conditions, Random Forest with its generally accurate performance was more erroneous by its higher RMSE (4.12 m^3/day). These results reveal that time-dependent models, including LSTM, are of special use in dynamic prediction of multiphase flow, and ensemble tree models, including Gradient Boosting, provide good baseline performance and predictive accuracy stability.

4.2. Classification Performance: Flow Regime Prediction

We classified flow regimes as slug, stratified, annular and bubbly. Table 2 depicts the performance of classification of RF, GB and a Transformer-based model.

Table 2: Classification Metrics for Flow Regime Prediction

Model	Accuracy (%)	Precision	Recall	F1-Score
Random Forest	91.2	0.90	0.91	0.90
Gradient Boosting	93.5	0.93	0.94	0.93
Transformer	95.1	0.94	0.95	0.95

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 2 Feb. 2026, pp: 338-344 www.ijaem.net ISSN: 2395-5252

The results of the classification show that all of the models were good at distinguishing between the flow regimes with an accuracy of more than 90%. Transformer model is the highest accuracy (95.1%) and F1-score (0.95), indicating its efficacy in the long-range temporal association of sequence pipeline data. Also powerful was the Gradient Boosting but with slightly lower metrics and good generalization across regimes. Random Forest did not exhibit the same accuracy and F1-score, but the difference between them was relatively small, especially compared with transitional flow regimes like slug to annular, indicating that it is less suited to detecting subtle changes in the pattern of the flow. In general, models based on the utilization of the temporal connections can make more justified and correct predictions of the flow regimes.

4.3. Robustness Testing

Gaussian noise with the input measurements was applied to test robustness (5% of standard deviation of features) to the measurements. Table 3 gives RMSE on flow rate prediction in the condition of noise.

Table 3: RMSE Under Noise (m^3/day)

Model	Clean Data	Noisy Data	ΔRMSE
Random Forest	4.12	5.05	+0.93
Gradient Boosting	3.79	4.31	+0.52
LSTM	3.65	4.02	+0.37

The robustness test indicates all the models were impacted by the addition of the Gaussian noise to the input features to some degree indicating the influence of the measurement uncertainties to the real pipeline data. LSTM showed the least growth in RMSE (+0.37 m^3/day), which demonstrates its better capacity to process the sequential data with lots of noise by means of temporal smoothing. Gradient Boosting had fairly constant predictions with a medium change increase (+0.52 m^3/day), whereas Random Forest was mostly affected (+0.93 m^3/day), indicating a lack of resistance to input noise. These findings point out that time-based models, such as LSTM, are more appropriate to strong multiphase flow forecasting in uncertain or noisy situations.

V. Conclusion and Recommendations

- **Conclusion**

This research created a standard test set and set of evaluation suites to test machine learning models in multiphase pipeline flow forecasting. The findings showed that temporal deep learning models especially the LSTM and Transformer models outperformed regular regression and classification models. In flow rate prediction, these models recorded high predictive accuracy with (R^2 =0.95) and RMSE values of between 3.65 m^3/ day and 3.79 m^3/ day. They demonstrated high accuracy, 95.1% and F1-score of 0.95 in the classification of flow regimes, and they are also highly robust to noisy and dynamically varying flow conditions.

Gradient Boosting Ensemble methods offered consistent and predictable baseline performance, which validates their application to nonlinear multiphase flow problems. Random Forest models however exhibited a relatively lesser resilience in transient and highly dynamic situations. Generally, the results suggest that standardized datasets, consistent evaluation procedures, and time modeling methods are necessary to secure reproducible, sound, and precise machine learning applications in a multiphase pipeline functioning.

- **Recommendations**

According to the findings, the future research needs to focus on the development of the benchmark

datasets to involve greater field data with the increased operating conditions. It is suggested to use physics-informed machine learning methods to improve generalization and explainability in pipeline settings that involve safety issues. Moreover, an extension of the current studies of advanced attention-based and hybrid deep learning architectures would enhance performance in highly transient flow conditions. Lastly, it is highly recommended that standardized evaluation frameworks be adopted between studies to allow fair comparison, reproducibility, and faster realization of stable ML solution to industrial pipeline monitoring and optimization.

REFERENCES

[1]. Alakbari, F. S., Ayoub, M. A., Awad, M. A., Ganat, T., Mohyaldinn, M. E., & Mahmood, S. M. (2025). A robust pressure drop prediction model in vertical multiphase flow: A machine learning approach. Scientific Reports, 15, Article 13420. https://doi.org/10.1038/s41598-025-96371-2 pure.kfupm.edu.sa

[2]. Alizadeh, R., & Others. (2021). Comparison of machine learning methods for multiphase flowrate prediction. In 2019 IEEE International Conference on Imaging Systems and Techniques (IST) (pp. 1–6). IEEE. https://doi.org/10.1109/IST48021.2019.9010450

[3]. Alizadeh, S., Mohammadi, A. and Rashtchian, D. (2021). Machine learning approaches for multiphase flow regime prediction in pipelines. *Journal of Petroleum Science and Engineering*, 203, 108605.

[4]. Chen, Y. and Li, X. (2023). Hybrid deep learning framework for transient multiphase flow prediction in pipelines. *Energy Reports*, 9, 2143–2156.

[5]. Gao, P. and Feng, Y. (2021). Data-driven modeling of pressure drop in gas–liquid multiphase flow systems. *Chemical Engineering Research and Design*, 171, 354–366.

[6]. Guo, H., Zhang, T. and Liu, Q. (2022). Comparative evaluation of machine learning models for multiphase flow classification. *Flow Measurement and Instrumentation*, 84, 102145.

[7]. Guo, L., Hu, D., Chen, W., Li, Y., Zhang, M., & Peng, L. (2021). CNNbased volume flow rate prediction of oil–gas–water threephase intermittent flow from multiple sensors. Sensors, 21(4), 1245. https://doi.org/10.3390/s21041245

[8]. Hafsa, N., Rushd, S., & Yousuf, H. (2023). Comparative performance of machinelearning and deeplearning algorithms in predicting gas–liquid flow regimes. Processes, 11(1), 177. https://doi.org/10.3390/pr11010177

[9]. Hosseini, M., Karimi, H. and Shafiei, A. (2022). Unsupervised learning for anomaly detection in multiphase pipelines. *Process Safety and Environmental Protection*, 160, 682–694.

[10]. Li, J., Hu, D., Chen, W., Li, Y., Zhang, M., & Peng, L. (2021). CNNbased volume flow rate prediction of oil–gas–water threephase intermittent flow from multiple sensors. Sensors, 21(4), 1245. https://doi.org/10.3390/s21041245 MDPI

[11]. Liu, Z., Wang, J. and Chen, D. (2022). Ensemble learning methods for prediction of liquid holdup in two-phase flow pipelines. *Applied Energy*, 314, 118932.

[12]. Mohammadi, A. and Rafiee, M. (2023). Gradient boosting models for multiphase flow pressure gradient estimation. *Petroleum Exploration and Development*, 50(2), 412–423.

[13]. Mohammadi, F., Ranjbar, A., & Kafi, M. (2023). Application of machine learning algorithms in classification of flow units of the Kazhdumi reservoir in one of the oil fields in southwest of Iran. Journal of Petroleum Exploration and Production Technology, 13, 1419–1434. https://doi.org/10.1007/s13202-023-01618-1

[14]. Rahman, F. and Ibrahim, S. (2023). Explainable artificial intelligence for flow regime prediction in oil–gas pipelines. *Expert Systems with Applications*, 213, 119214.

[15]. Saparbayeva, N., Balakin, B. V., Struchalin, P. G., Rahman, T., & Alyaev, S. (2024). Application of machine learning to predict blockage in multiphase flow. Computation, 12(4), 67. https://doi.org/10.3390/computation12040067 MDPI

[16]. Sun, Y., Zhao, L. and Huang, R. (2024). Transformer-based time-series modeling for industrial flow systems. *Engineering Applications of Artificial Intelligence*, 124, 106512.

[17]. Wang, L., Xu, Z. and Peng, H. (2021). LSTM-based prediction of slug flow dynamics in multiphase pipelines. *Journal of Natural Gas Science and Engineering*, 90, 103925.

[18]. Zhang, H., Yang, Y., & Others. (2020). A novel CNN modeling algorithm for the instantaneous flow rate measurement of gas–liquid multiphase flow. In Proceedings of ICMLC 2020. https://yyysjz1997.github.io/Files/Zhang_2020_ICMLC.pdf

[19]. Zhang, X., Li, Y. and Zhou, J. (2020). Deep convolutional neural networks for gas–liquid flow regime classification. *Measurement*, 162, 107884.