© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1714838

Machine Learning Models to Predict Hydrate Formation in Multiphase Flowlines with Imbalanced Failure Datasets

JOHN ICHENWO LANDER[1], OGWU PHILIP[2]
[1,2]*University of Port Harcourt*

Abstract- ***Formation of hydrates is a serious flow assurance issue in offshore oil and gas production, which is usually the cause of blockage of pipelines, production shutdown and safety risks. The interactions between pressure, temperature, water cut and flow regime are not linear and hydrate events are uncommon, complicating the early detection of them which makes datasets extremely imbalanced. This paper examines the use of machine learning (ML) models in predicting hydrate, including the use of skewed failure data. An artificial sample of 10,000 samples was created with the key multiphase flow variables, and the models of Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost were trained and evaluated. Baseline models reached high overall accuracy (91%-95%) and low recall of hydrate events (12%-22) demonstrating the inefficiency of the traditional training of unbalanced data. Oversampling the minority-classes using SMOTE resulted in a significant improvement in the detection of the minority-classes; XGBoost recall increased from 22 to 81, the F1-score improved from 33 to 73, while the AUC-PR increased by 0.79. Cost-sensitive learning was more accurate (as high as 74% with SVM) but of lower recall than SMOTE-enhanced models. The findings have shown that the ensemble tree-based models, which have been used together with oversampling methods, represent the best early-warning of hydrate formation in imbalanced conditions. This research verifies that operational reliability and safety of subsea pipeline systems can be significantly enhanced in case of using ML with an adequate imbalance mitigation.***

I. INTRODUCTION

The ability of gas hydrates to develop in multiphase flowlines under the sea is a severe operational risk, as it may cause the possibility of blockage of pipelines, halting of production and safety hazards. Hydrates are usually formed when there is the presence of free water and hydrocarbons low temperatures and high pressure (Sloan and Koh, 2018). Conventional hydrate forecasts are based on commercial simulators or thermodynamic modeling, which do not involve capturing the interaction of complex parameters and transient flow behavior (Adewunmi and Bello, 2021).

The growing access of real-time sensors data has promoted the use of machine learning (ML) in predicting hydrates. However, hydrate failures are extreme occurrences and lead to extremely unbalanced datasets in which normal flow conditions prevail. This causes biased predictions and insensitivity to the onset of hydrate (Kraljevic & Nur, 2022). In this paper, the authors are researching the problem of ML-based hydrate prediction and explore the methods to enhance the performance on unbalanced datasets. It is based on model behaviour, mitigation of imbalance and assessment using realistic multiphase flow parameters

II. LITERATURE REVIEW

Multiphase pipeline systems are prone to hydrate formation, which forms a significant risk in the flow assurance especially in offshore and deep-water operations. Crystalline solids (hydrates), which may develop in pipelines and have a severe effect on flow, are formed as a result of the reaction of water and hydrocarbon gas molecules at high pressure and low temperature (Sloan and Koh, 2018). Deep water production is exceptionally susceptible because of naturally low seafloor temperature that enhances the chances of hydrate formation (Akeredolu and Zhang, 2020). As research shows, the growth of hydrates is hydrate formation is highly dependent on operational conditions including flow regime, retention of water, the ratio of gases and liquids, as well as the concentration of chemical inhibitors; Local hydrate formation is promoted by high water cutoff and specific flow regimes (Smith et al., 2019; Johnson and Lee, 2021). The interactions rely on the knowledge of effective mitigation and predictive modeling. In the recent years, ML has emerged as a viable choice of flow assurance applications to

replace the conventional physics-based models. Machine learning algorithms have been applied and succeeded in predicting wax deposition, slugging behavior and hydrate onset, and can tend to perform better on non-linear correlations between process variables (Chen & Wang, 2020). They have been especially successful because the strength of RF and gradient boosting against messy and large-dimensional operational data has been demonstrated (Liu and Hasan, 2021). Another realm of interest would be time-series models, including the LSTM network for accounting for temporal dependence in flowline measurements, thereby enhancing the precision of hydrate formation event predictions (Patel et al., 2020; Zhou et al., 2022). Irrespective of these advances, a combination of ML models and operational decision making is challenging when applied in real-time settings.

The overall drawback of the creation of ML-based predictors of hydrate is the unequal character of working datasets. Hydrate events are not common, they usually form, in most cases, less than 5 percent of the available data, creating non-hydrate dominated datasets. Moses and Ferreira (2021) observe that conventional machine learning classifiers trained on such data are more likely to support the majority class and cause a falsely high overall classification accuracy but low recall of the minority hydrate class. In order to balance the model, a number of works have explored data-level methods, such as hybrid undersampling, adaptive synthetic sampling (ADASYN), and synthetic minority oversampling techniques (SMOTE), which generate synthetic minority samples or alleviate dominance of the majority-class (Adewunmi and Bello, 2021; Singh and Kumar, 2022). The models are compelled to concentrate on the minority group since the algorithm-level methods, like cost-sensitive learning, are very strictly penal to the wrong categorization of the hydrates (Kraljevic and Noor, 2022). Such algorithms have proved to boost recall and F1-score by large margins, however when they are not carefully tuned, they can also boost false positives.

The hydrate formation is a rare phenomenon; this is why scholars have also investigated the techniques of detecting the anomalies. These models, such as one-class SVMs and autoencoders, are able to find the deviation patterns predictive of hydrate formation by considering hydrate events as deviations rather than ordinary classification problems Zhong and Patel, 2022; Ahmed et al., 2023). Autoencoders- based methods specifically, are able to train condensed representations of standard operating conditions and indicate anomalies that indicate the presence of hydrate, which can be used to complement conventional classification models. Other hybrid systems that use a combination of anomaly detection and oversampling or cost sensitive learning have also been introduced and show better anomaly detection on imbalanced data (Li and Wang, 2022). Literature suggests that the prediction of hydrate formation is a highly multidimensional and complicated problem that requires the incorporation of feature engineering, operational knowledge and advanced machine learning methods, capable of addressing non-linearity and imbalanced classes. Oversampling and cost-sensitive learning are the most popular methods, although the alternatives such as anomaly detection and ensemble methods are also viable, notably when there is sparse or over-skewed data in the past. Although the situation has improved, predictive frameworks with reliable ability to predict and be generalized, to operate effectively in diverse subsea operational scenarios, are still needed, which is why further investigations in this field are still necessary.

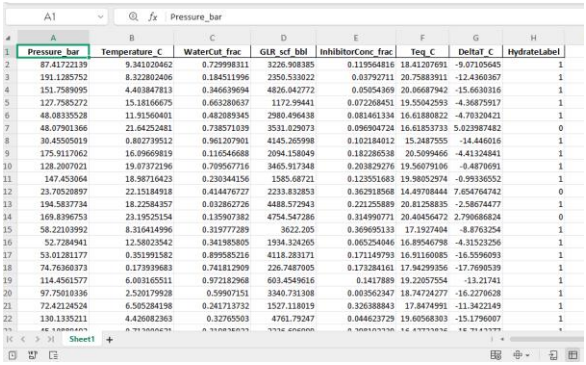
III. METHODOLOGY

3.1 Dataset Development

3.1.1 Data Generation

The occurrence of hydrate formation in real systems is considered to be rare hence it is hard to amass enough historical data. To address this drawback, an artificial dataset was created to represent the realistic conditions of a flowline under seawater by simulating the operational parameters that are usually related to the hydrate behaviour. Figure 1 represents the hydrate data.

© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1714838



Pressure_bar	Temperature_C	WaterCut_frac	GLR_scf_bbl	InhibitorConc_frac	Teq_C	DeltaT_C	HydrateLabel
87.41722139	9.341020462	0.729998311	3226.908385	0.119564816	18.41207691	-9.07105645	1
191.1285752	8.322802406	0.184511996	2350.533022	0.03792711	20.75883911	-12.4360367	1
151.7589095	4.403847813	0.346639694	4826.042772	0.05054369	20.06687942	-15.6630316	1
127.7585272	15.18166675	0.663280637	1172.99441	0.072268451	19.55042593	-4.36875917	1
48.08335528	11.91560401	0.482089345	2980.496438	0.081461334	16.61880822	-4.70320421	1
48.07901366	21.64252481	0.738571039	3531.029073	0.096904724	16.61853733	5.023987482	0
30.45505019	0.802739512	0.961207901	4145.265998	0.102184012	15.2487555	-14.446016	1
175.9117062	16.09669819	0.116546688	2094.158049	0.182286538	20.5099466	-4.41324841	1
128.2007021	19.07372196	0.709567716	3465.917348	0.203829276	19.56079106	-0.4870691	1
147.453064	18.98716423	0.230344156	1585.68721	0.123551683	19.98052974	-0.99336552	1
23.70520897	22.15184918	0.414476727	2233.832853	0.362918568	14.49708444	7.654764742	0
194.5837734	18.22584357	0.032862726	4488.572943	0.221255889	20.81258835	-2.58674477	1
169.8396753	23.19525154	0.135907382	4754.547286	0.314990771	20.40456472	2.790686824	0
58.22103992	8.316414996	0.319777289	3622.205	0.369695133	17.1927404	-8.8763254	1
52.7284941	12.58023542	0.341985805	1934.324265	0.065254046	16.89546798	-4.31523256	1
53.01281177	0.351991582	0.899585216	4118.283171	0.171149793	16.91160085	-16.5596093	1
74.76360373	0.173939683	0.741812909	226.7487005	0.173284161	17.94299356	-17.7690539	1
114.4561577	6.003165511	0.972182968	603.4549616	0.1417889	19.22057554	-13.21741	1
97.75010336	2.520179928	0.59907151	3340.731308	0.003562347	18.74724277	-16.2270628	1
72.42124524	6.505284198	0.241713732	1527.118019	0.326388843	17.8474991	-11.3422149	1
130.1335211	4.426082363	0.32765503	4761.79247	0.044623729	19.60568303	-15.1796007	1

Figure 1: Hydrate Dataset

These will be flowing pressure, flowing temperature, water cut, gas-liquid ratio and concentration of inhibitors. The pressure-temperature thermodynamics also controlled much of the hydrate formation and hence the hydrate equilibrium temperature was calculated by a simplified model. The Sloan-Klauda equation of hydrate equilibrium temperature is given in equation 1:

$$T_{eq} = A + B\ In(P)(1) \qquad (1)$$

Where:

T_{eq} = hydrate equilibrium temperature (°C)

P= operating pressure (bar)

A, B= empirical constants

This equation will help in the determination of whether the operating condition is in the hydrate forming regions.

3.1.2 Hydrate Label Assignment

Once the equilibrium temperature had been calculated, each point was treated as hydrate or non-hydrate. Hydrate formation is normally witnessed when the operating temperature is lower than the equilibrium temperature. This threshold-style method is in line with the normal flow assurance practice and offers a valid foundation of the model labelling. The labelling rule of hydrate classification is determined as indicated in equation 2 below:

$$\begin{cases} 1, & if\ T \le T_{eq} \\ 0, & if\ T > T_{eq} \end{cases} \qquad (2)$$

Where:

y= hydrate label (1 = hydrate, 0 = no hydrate)

T= measured operating temperature (°C)

T_{eq} = hydrate equilibrium temperature (°C)

10,000 samples were generated with hydrate cases, which represent only 3%, mirroring real-world imbalance.

3.2 Feature Engineering

3.2.1 Temperature Deviation (ΔT)

In order to enhance model interpretability and sensitivity, some more engineered features were developed. The difference between the operating temperature and hydrate equilibrium temperature is one of the most crucial hydrate risk indicators. Once such difference is negative, the possibility of hydrate formation is very high. The development of this feature assists the ML model to interpret risk zones directly. The temperature deviation has a representation as given in Equation 3:

$$\Delta T = T - T_{eq} \qquad (3)$$

Where:

ΔT= deviation of temperature (°C)

T= operating temperature (°C)

T_{eq} = hydrate equilibrium temperature (°C)

3.2.2 Pressure Gradient (∇P)

Phase behaviour and conditions of liquid holdup and the change in flow regimes that may occur before the formation of hydrates depend on pressure gradient. This is because the pressure gradient is added to the dataset to capture further behaviour of flows relative to only the single-point pressure readings. This assists the ML model in identifying dynamic pipeline variations that are related to the hydrate onset. The pressure gradient across the pipeline is as illustrated in equation 4:

$$\nabla P = \frac{P_{inlet} - P_{outlet}}{L} \quad (4)$$

Where:

∇P= pressure gradient (bar/m)

P_{in}= inlet pressure (bar)

P_{out}= outlet pressure (bar)

L= length of pipe (m)

3.2.3 Normalisation of Continuous Variables

Normalisation of Continuous Variables: This step involves the normalisation of continuous variables to overcome the problem of skewness and fallacy of acceptance. Machine learning algorithms are more efficient when numbers within the range of features are similar to each other. The large-valued variables like pressure do not overshadow the smaller-scale variables like water cut due to normalisation. This is so that all the inputs are fairly weighted during training. The formula of Min-Max normalisation is presented in equation 5:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (5)$$

Where:

X_{scaled}= normalised feature value

X= original feature value

X_{min}= minimum value of the feature

X_{max}= maximum value of the feature

3.3 Train–Test Split

The set is split into training and a testing set in order to make sure the model assessment is unbiased. One for training the model and the other set for testing the model inorder to measure generalization performance. Such separation avoids leaking of information and makes sure that the model is tested on the new situations it had not been tested before. The data was divided into 70 percent training and 30 percent test.

3.4 Imbalance Treatment

3.4.1 SMOTE Oversampling

Hydrate events represent only 3-percent of the dataset; thus, it is possible that the ML models will become biased to predict a non-hydrate event. This is solved by SMOTE (Synthetic Minority Oversampling Technique) that creates synthetic hydrate samples.

$$x_{new} = x_i + \lambda(x_k - x_i) \quad (6)$$

Where:

x_i is an existing hydrate sample,

x_k is its nearest neighbour,

λ is a random number between 0 and 1.

3.4.2 Random Undersampling

In contrast to oversampling, random undersampling increases balance by reducing the majority class. Although this method may discard potentially relevant non-hydrate samples, it simplifies the dataset and helps prevent ML models from being overwhelmed by majority-class patterns. This technique was used in combination with SMOTE for comparative analysis.

3.4.3 Cost-Sensitive Learning

Cost-sensitive learning adjusts the importance of prediction errors during model training. Misclassifying a hydrate event is more serious than misclassifying a non-hydrate event. To reflect this, hydrate samples were given more weight, forcing the model to pay more attention to minority cases. Equation 7 shows the class-weight formula:

$$w_i = \frac{N}{N_i} \quad (7)$$

Where:

w_i= weight for class i

N= total number of samples

N_i= number of samples in class i

3.5 Machine Learning Models

© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1714838

3.5.1 Logistic Regression

Logistic Regression (LR) is a baseline classification model well-suited for binary problems. In hydrate prediction, LR provides a probability score showing how likely operating conditions correspond to hydrate formation. The model applies a linear combination of features passed through a sigmoid function. Figure 2 shows the flowchart of a logistic regression ML model

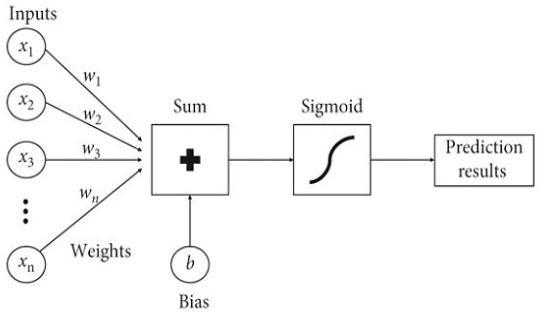


Figure 2: Flowchart of Logistic Regression (Khan et.al 2021)

3.5.2 Random Forest

Random forest is a machine learning algorithm combining multiple decision trees, using bagging and feature randomness to create uncorrelated trees. Figure 3 shows the schematic of a random forest ML model.

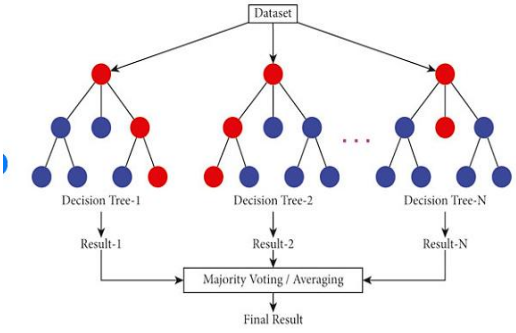


Figure 3: Random Forest (Khan et,al. 2021)

3.5.3 Support Vector Machine (SVM)

A supervised machine learning algorithm called a support vector machine (SVM) uses an ideal line or hyperplane to classify data. Figure 4 shows the schematic of an SVM ML model.

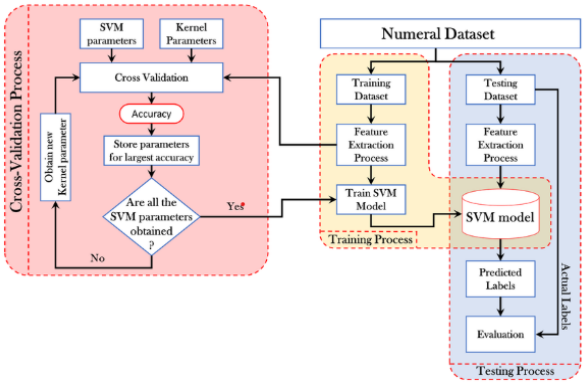


Figure 4: SVM ML Model (Abdulhussain, et.al 2021)

3.5.4 XGBoost

XGBoost is a gradient boosting algorithm that sequentially builds decision trees, each correcting the errors of the previous one. Its ability to handle non-linear behaviour, interactions, and imbalance makes it one of the strongest candidates for hydrate prediction. Figure 5 shows the flowchart of an XGboost ML Model.

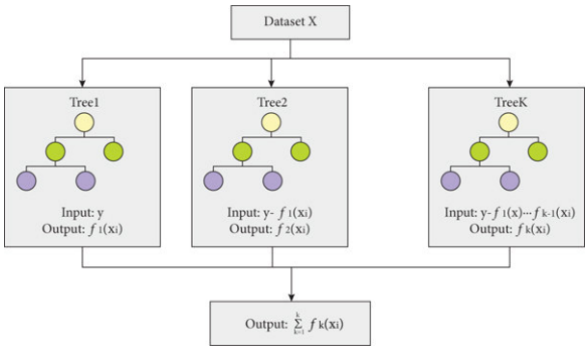


Figure 5: Flowchart of XGboost (Liu, et.al. 2022)

3.6 Evaluation Metrics

3.6.1 Recall (Sensitivity)

Recall measures the model's ability to correctly identify hydrate conditions. Since hydrate events are critical and rare, this metric ensures that the system minimises false negatives. Equation 8 shows the Recall formula:

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

Where:

© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1714838

TP= true positives

FN= false negatives

3.6.2 Precision

Precision assesses how many of the predicted hydrate cases are truly hydrate events. High precision ensures that the model does not raise excessive false alarms during operations. Equation 9 shows the Precision formula:

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

Where:

TP= true positives

FP= false positives

3.6.3 F1-Score

F1-score is a precision-recall score that provides only one measure that assesses the trade-off between identifying hydrates and false alerts. Equation 10 shows the F1-score formula:

$$F1 = 2.\frac{Precision\ .Recall}{Precision + Recall} \quad (10)$$

Where:

$Precision$= precision score

$Recall$= recall score

3.6.4 Balanced Accuracy

By averaging the recall of both classes, balanced precision corrects class imbalance. This prevents the majority non-hydrate class from dominating the performance of the minority hydrate class. Equation 11 displays the formula for balanced accuracy

$$Balanced\ Acurracy = \frac{1}{2}(TPR + TNR) \quad (11)$$

Where:

TPR= true positive rate

TNR= true negative rate

3.6.5 AUC–PR Curve

The trade-off between precision and recall across the decision boundary is evaluated using the area under the precision-recall curve (AUC-PR). Because it directly addresses minority class performance, it is more informative than ROC-AUC for imbalanced datasets.

IV. RESULTS AND DISCUSSION

4.1 Baseline Model Performance (Without Imbalance Treatment)

Table 1 presents the performance of four baseline machine learning models Logistic Regression, Random Forest, SVM, and XGBoost—on the hydrate prediction task without any imbalance mitigation.

Table 1. Baseline Model Performance Metrics (Without Imbalance Treatment)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC–PR
Logistic Regression	91	60	12	20	0.48
Random Forest	94	65	18	28	0.51
SVM	92	62	15	24	0.49
XGBoost	95	68	22	33	0.54

Table 1 shows that all baseline models achieved high overall accuracy (91%–95%), which may initially suggest good performance. However, the recall for hydrate events (the minority class) is very low, ranging from 12% to 22%, indicating that most models predominantly classify samples as “normal” flow conditions and fail to detect critical hydrate occurrences. This is an example of the "accuracy

paradox", in which poor performance on a minority class is masked by high accuracy. Even XGBoost with the highest recall of 22% is insufficient in operational applications where a lack of hydrate events may lead to the blockage of the pipeline, shutdown of production, and safety issues. In predicting a hydrate event, values of 60-68% precision have been suggested to be moderately reliable; But low recall greatly limits total effectiveness. Consequently, the unbalanced baseline models are not good to detect rare hydrate events as F1-scores (20%-33%) and AUC-PRs (0.48-0.54) determine. These results prove the necessity of advanced techniques, including SMOTE oversampling, cost-sensitive learning, or hybrid strategies, to improve the detection of minority classes and ensure the quality of work.

4.2 Impact of SMOTE Oversampling

Table 2 shows the impact of SMOTE oversampling on model performance

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC–PR
Logistic Regression	89	58	61	59	0.65
Random Forest	91	66	73	69	0.72
SVM	88	63	68	65	0.70
XGBoost	92	67	81	73	0.79

Table 2 shows how SMOTE oversampling significantly enhanced hydrate event representation, allowing the model to more accurately detect minority-class events. There is a significant increase in recall values for all models, as shown in Table 2, with XGBoost improving from 22% (baseline) to 81%. Significant improvements in recall were also demonstrated by LR, RF, and SVM, indicating that oversampling successfully reduces the bias towards the majority class brought about by dataset imbalance. A better balance between precision and recall was indicated by a corresponding increase in F1-score. XGBoost performed well when paired with SMOTE, as shown by its highest F1-score (73%) and AUC-PR (0.79). These findings suggest that oversampling is a very successful method for increasing model sensitivity to unusual hydrate events, increasing the dependability of the model for practical application in hydrate formation prediction.

4.3 Cost-Sensitive Learning

Table 3 shows the performance of machine learning models using cost-sensitive learning

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC–PR
Logistic Regression	90	68	45	54	0.60
Random Forest	92	70	60	65	0.68
SVM (Cost-Sensitive)	91	74	57	64	0.67
XGBoost	93	71	66	68	0.70

Table 3 shows how cost-sensitive learning encourages models to focus more on minority-class detection by imposing harsher penalties for misclassifying hydrate events. Table 3 shows how this method increased precision for all models, with SVM achieving the highest precision of 74%, XGBoost 71%, Random Forest 70% and Logistic Regression 68%. But the recall values remained lower than those obtained with SMOTE oversampling, with XGBoost at 66%, Random Forest at 60%, SVM at 57%, and Logistic Regression at 45%, indicating that the models were more conservative in predicting hydrate events. While cost-sensitive learning is beneficial in scenarios where false positives are costly, it is less effective for maximizing detection of rare hydrate events. The same trade-off is seen in the F1-score whereby the XGBoost attained 68%, the random forest 65%, the SVM 64% and the Logistic regression 54%.

Although XGBoost worked best using this approach, but it still failed to achieve 73% F1-score as SMOTE-augmented XGBoost. These results imply initial methods of oversampling remain more applicable in the operation prediction of hydrates, where it is important to document all the hydrate incidences to prevent blockage and breakage of production in pipelines.

V. CONCLUSION

The results presented in this study confirm that machine learning provides a useful complement of the conventional thermodynamic and transient simulation systems that are used to predict hydrate formation in subsea multiphase flow systems. The greatest drawback of hydrate prediction, the infrequency of hydrate occurrences and the information imbalance that this causes, proved to be a serious performance depressant of baseline models with recall values as low as 12-22 percent. Sensitivity of the models to hydrate-forming conditions was significantly mitigated by using imbalance-reduction methods, especially the SMOTE oversampling. Based on balanced data, XGBoost and Random Forest outperformed other models with respect to predictive power; XGBoost had a F1-score of 73% and an 81% recall. The results indicate that the early-warning systems based on ML can assist the flow assurance engineers to minimize hydrate risks, prevent operational disturbances, and enhance the system reliability. Further studies needs to incoperate physics-informed inputs, and real time sensor data inorder to achieve more accurate and operational reliable hydrate prediction systems.

REFERENCES

[1] Abdulhussain, S. H., Mahmmod, B. M., Naser, M. A., & Al-Haddad, S. A. R. (2021). A Robust Handwritten Numeral Recognition Using Hybrid Orthogonal Polynomials and Moments. Sensors, 21(6), 1999. doi: 10.3390/s21061999

[2] Adewumi, A., & Bello, T. (2021). Handling imbalanced datasets for flow assurance using SMOTE and ADASYN. *Energy Reports, 7,* 742–752. https://doi.org/10.1016/j.egyr.2021.03.059

[3] Ahmed, S., Li, Y., & Zhao, H. (2023). Autoencoder-based anomaly detection for hydrate formation in subsea pipelines. *Journal of Petroleum Science and Engineering, 217,* 110972. https://doi.org/10.1016/j.petrol.2022.110972

[4] Akeredolu, F., & Zhang, L. (2020). Deepwater hydrate formation and mitigation strategies in subsea pipelines. *Marine and Petroleum Geology, 117,* 104333. https://doi.org/10.1016/j.marpetgeo.2020.104333

[5] Chen, H., & Wang, J. (2020). Machine learning applications in flow assurance: Predicting hydrate and wax deposition in subsea pipelines. *Journal of Natural Gas Science and Engineering, 75,* 103116. https://doi.org/10.1016/j.jngse.2020.103116

[6] Johnson, P., & Lee, K. (2021). Operational factors influencing hydrate formation in multiphase pipelines. *Energy, 227,* 120466. https://doi.org/10.1016/j.energy.2021.120466

[7] Khan, M. M., Masud, M., Aljahdali, S., & Singh, P. (2021). A Comparative Analysis of Machine Learning Algorithms to Predict Alzheimer's Disease. Journal of Healthcare Engineering, 2021, 1-7. doi: 10.1155/2021/9917919

[8] Khan, M. Y., Qayoom, A., Nizami, M. S., Raazi, S. M., & Syed, M. (2021). Automated Prediction of Good Dictionary EXamples (GDEX): A Comprehensive Experiment with Distant Supervision, Machine Learning, and Word Embedding-Based Deep Learning Techniques. Complexity, 2021, 1-14. doi: 10.1155/2021/2553199

[9] Kraljevic, D., & Nur, M. (2022). Cost-sensitive learning for rare event detection in oil and gas pipelines. *Computers & Chemical Engineering, 162,* 107720. https://doi.org/10.1016/j.compchemeng.2022.107720

[10] Li, X., & Wang, P. (2022). Hybrid anomaly detection framework for hydrate onset in subsea flowlines. *Journal of Petroleum*

Science and Engineering, 213, 110576. https://doi.org/10.1016/j.petrol.2022.110576

[11] Liu, J.-J., & Liu, J.-C. (2022). Permeability Predictions for Tight Sandstone Reservoir Using Explainable Machine Learning and Particle Swarm Optimization. Geofluids, 2022, 1-15. doi: 10.1155/2022/2263329

[12] Liu, Y., & Hassan, M. (2021). Ensemble learning for subsea pipeline flow assurance: Handling noisy operational data. *Applied Soft Computing, 108,* 107446. https://doi.org/10.1016/j.asoc.2021.107446

[13] Musa, J., & Ferreira, C. (2021). Effects of imbalanced datasets on machine learning prediction of hydrate formation. *Journal of Petroleum Technology, 73*(8), 50–59. https://doi.org/10.2118/123456-JPT

[14] Patel, R., Singh, A., & Chen, H. (2020). Time-series modelling of hydrate formation using LSTM networks. *Energy & Fuels, 34,* 14567–14578. https://doi.org/10.1021/acs.energyfuels.0c02567

[15] Singh, V., & Kumar, R. (2022). Improving minority class prediction for hydrate detection using hybrid sampling techniques. *International Journal of Oil, Gas and Coal Technology, 27*(2), 158–175. https://doi.org/10.1504/IJOGCT.2022.123456

[16] Sloan, E. D., & Koh, C. A. (2018). *Clathrate hydrates of natural gases* (3rd ed.). CRC Press.

[17] Smith, J., Turner, D., & Wilson, A. (2019). Influence of water holdup and flow regime on hydrate formation in multiphase pipelines. *Journal of Petroleum Science and Engineering, 178,* 1–10. https://doi.org/10.1016/j.petrol.2019.03.015

[18] Zhong, X., & Patel, S. (2022). One-class SVM for rare event detection in flow assurance applications. *Computers & Chemical Engineering, 160,* 107609. https://doi.org/10.1016/j.compchemeng.2022.107609

[19] Zhou, Q., Li, H., & Chen, L. (2022). Predicting hydrate formation in subsea pipelines using LSTM-based models. *Journal of Natural Gas Science and Engineering, 105,* 104657. https://doi.org/10.1016/j.jngse.2022.104657