2025 International Conference on Data Science, Agents, and Artificial Intelligence (ICDSAAI 2025) 28-29 March 2025, Chennai, India

Classification of Multiphase Flow Regime in a Horizontal Pipe Using Supervised Machine Learning and Bayesian Optimization

Abhilash Ravichandran
Department of Chemical Engineering
NIT Andhra Pradesh
Tadepalligudem-534101, India
221101@student.nitandhra.ac.in

Namburi Pooja Sai Venkata Sri Harika
Department of Chemical Engineering
NIT Andhra Pradesh
Tadepalligudem-534101, India
222123@student.nitandhra.ac.in

Sai Manikiran Garimella
Department of Chemical Engineering
NIT Andhra Pradesh
Tadepalligudem-534101, India
gmanikiran@gmail.com

Mohan Anand
Department of Chemical Engineering
IIT Hyderabad
Sangareddy-502285, Telangana, India
anandm@che.iith.ac.in

***Abstract*—Multiphase flow refers to a fluid flow consisting of more than one phase. Conventional/traditional methods rely on empirical correlations and flow maps of the flow regimes that show overlapping characteristics. Machine learning (ML) has gained traction for multiphase flow regime classification, offering improved accuracy and adaptability to complex flow conditions. This paper presents the supervised ML approach to classify the three-phase flow regimes in horizontal pipes. The superficial gas and liquid velocities and water cuts are utilized as input features for flow regime classification. ML classification models, such as Bayesian Classifier and K-Nearest Neighbors, and boosting models, such as XGBoost, CatBoost, and LightGBM, are employed and evaluated. Bayesian optimization is used to enhance model accuracy and robustness by hyperparameter tuning. The selected supervised ML models provided an accurate and efficient classification of multiphase flow regimes when enhanced using Bayesian optimization. The optimized ML models significantly improved the classification accuracy and the F1 score, making them practical for real-time monitoring and control applications.**

***Index Terms*—Multiphase flow, Machine Learning, Gaussian Naive Bayes, K-Nearest Neighbors, Boosting models, Bayesian optimization, F1-Score.**

I. Introduction

Multiphase flow is the simultaneous flow of multiple distinct phases, such as gas, liquid, or solid, showing complex flow patterns that are difficult to classify accurately. The classification of multiphase flow regimes plays an important role in industrial operations such as oil and gas extraction, chemical manufacturing, and energy production [1]. Three-phase flows involving air, water, and oil are ubiquitous, as they are frequently encountered in pipeline systems where different phases interact and influence each other's flow behavior. Multiphase flows involve the movement of different gas-liquid, liquid-liquid, or gas-solid phases and exhibit complex flow patterns such as stratified, bubbly, slug, churn, and annular flow. The flow regimes significantly influence the pressure drops, heat transfer rates, and operational efficiency, making their accurate identification necessary for process optimization and risk mitigation [2].

The precise identification of the multiphase flow regime is essential for the optimization of operational efficiency, safety, and process control [3]. However, the overlapping characteristics of the flow regimes make conventional classification methods challenging. Traditional methods rely on empirical correlations and flow maps to classify the flow regimes [4]. For instance, Mandhane and co-workers tested the different flow pattern maps for gas-liquid two-phase flow in horizontal pipes using traditional approaches, such as empirical correlations and flow regime maps [5]. Later, Zhang and Sarica developed the unified model to predict the flow behavior of gas/oil/water three-phase pipe flow, differentiating between stratified & dispersed oil/water flows [6]. Additionally, Petalas and co-workers proposed mechanistic models and introduced empirical correlations for interfacial friction and liquid fraction entrainment, which enhance multiphase flow calculations [7]. Salgado and co-workers demonstrated gamma-ray-based Artificial Neural Networks (ANN) methods for flow regime estimation and volume fraction prediction in multiphase flows [8]. Later, ML gained traction for multiphase flow regime classification, offering improved accuracy and adaptability to complex flow conditions. In recent years, Supervised ML models emerged as powerful alternatives to address the traditional approach challenges, utilizing annotated training data for classification [9]. Al-Naser and co-workers developed an ANN model to detect the flow patterns in horizontal pipes and achieved 97% accuracy in correctly classifying flow patterns across a wide range of flow conditions [10]. Despite their merits, these models face challenges, including data scarcity, data imbalance, overfitting risks, and the computational demands of real-time applications. Recently, Gradient-boosting

979-8-3315-3755-5/25/$31.00

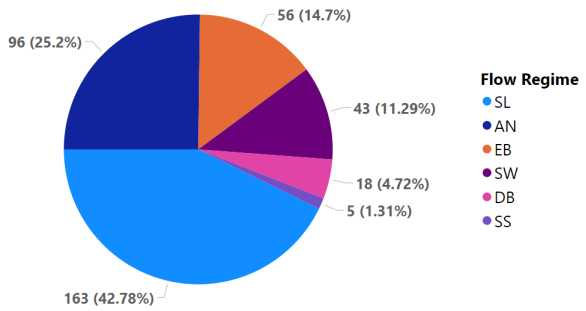

Fig. 1: Flow regime distribution: Percentage contribution by regime [13]

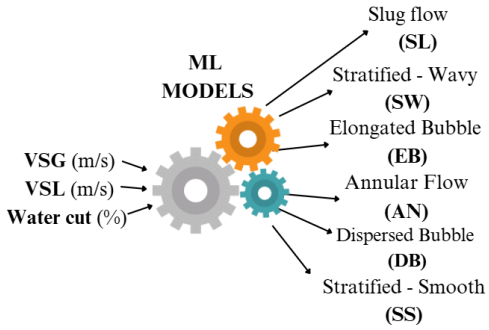

Fig. 2: Flow regime identification using ML models with input variables

algorithms, such as XGBoost, gained attention due to their efficiency and ability to handle diverse datasets. Wang and co-workers presented the Bayesian-optimized XGBoost model to predict the oil-water two-phase flow regimes and achieved better accuracy (93.8%) compared to traditional XGBoost (75%) [11]. Al Dogail and co-workers presented the dimensionless artificial intelligence model to predict the flow pattern in horizontal pipes and noted the improvement in accuracy using dimensionless parameters such as Reynolds and Weber numbers and extended their applicability to various operational conditions [12]. This paper studies the prediction of multiphase flow regimes in horizontal pipes carrying water, oil, and air phases using supervised ML approaches. The water cut, superficial gas and liquid velocity [13] are key input properties, and the six flow regimes are the potential outputs.

This paper is structured as follows: The dataset is gathered and fed into various models. These models are validated using the cross-validation approach outlined in Section II. The confusion matrix of six ML Models and a comparison of their training time, prediction time, and F1 scores are reported in section III. The results are summarized in section IV.

II. Methodology

A. Dataset

The dataset of flow regimes used for the study is collected from [13]. The dataset has four columns and 381 rows. The feature columns are superficial gas velocity, superficial liquid velocity, and water cut. The flow pattern is the target column. The flow regime distribution of the dataset collected from [13] is shown in Fig. 1.

The dataset exhibited a significant class imbalance, as Dispersed Bubble (DB) and Stratified-Smooth (SS) flow regimes are under-represented compared to the other regimes (SL, SW, EB, and AN flows) as shown in Fig. 1. ML models are biased toward majority classes, leading to suboptimal performance in the minority classes. The class weight adjustment technique is employed to overcome the class imbalance. The algorithm is incentivized by incorporating class weights to treat all classes more equitably and improve its ability to generalize across the entire dataset during the training of the model.

Class Weight Calculation: The class weights are calculated based on the frequency of occurrence of each flow regime to address the class imbalance in the multiphase flow dataset. The weight for each class w_i is given by:

$$w_i = \frac{N}{k \cdot n_i},$$

where w_i is the weight given to class i, N represents the total number of samples in the dataset, n_i denotes the number of samples in class i (frequency of occurrence of the flow regime), and k is the total number of classes in the dataset (number of distinct flow regimes) [14].

Six different machine learning models are used to classify flow regimes into six different outputs (flow regimes). The three input variables (superficial liquid velocity, superficial gas velocity, and water cut) and six output flow regimes (SL, SW, EB, AN, DB, and SS) that undergo prediction using ML models are displayed in Fig. 2. The selected supervised ML models, along with their mathematical frameworks, are described in the following section $II(B)$.

B. ML Models and Their Mathematical Frameworks

1) K-Nearest Neighbors (KNN): KNN algorithm determines a data point's label by considering the most common class among its closest neighbors in the feature space [15].

2) Naive Bayes: Naïve Bayes is a probability-based model built on Bayes' Theorem. Given the class, it is assumed that the features are independent of each other [16].

The posterior probability of a class y based on the features $x_1, x_2, \ldots, x_n$ is given by

$$P(y \mid x_1, x_2, \ldots, x_n) = \frac{P(x_1, x_2, \ldots, x_n \mid y)P(y)}{P(x_1, x_2, \ldots, x_n)}, \quad (1)$$

where P(y) is the prior probability of class y, $P(x_1, x_2, \ldots, x_n \mid y)$ is the likelihood of the features given y, $P(x_1, x_2, \ldots, x_n)$ is the evidence (constant for all classes).

The class-weight parameter adjusts the priors $P(y)$ to ensure fair representation of minority classes.

3) XGBoost: XGBoost is an advanced boosting algorithm that leverages decision trees to iteratively minimize a loss function.

The prediction after the m^{th} tree is given as

$$\hat{y}_i^{(m)} = \hat{y}_i^{(m-1)} + \eta f_m(x_i), \tag{2}$$

where η is the learning rate, and f_m is the m^{th} decision tree.

The cost function (which includes a loss term and a regularization term) is given by

$$\mathcal{L} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \Omega(f). \tag{3}$$

The $scale\_pos\_weight$ parameter is used and tuned to assign higher weights to the minority class [17].

4) CatBoost: CatBoost is a gradient-boosting algorithm that efficiently handles categorical data, using ordered boosting to prevent overfitting. CatBoost minimizes the cross-entropy loss for multi-class classification. Class weights are explicitly specified using $class\_weights$ parameter to ensure balanced training [18].

5) LightGBM: LightGBM is a gradient-boosting algorithm (designed for speed and memory efficiency) that uses histogram-based learning to accelerate computing.

6) Hybrid Stacking Model: The Hybrid Stacking model or stack model combines the predictions of LightGBM, XGBoost, and CatBoost (base learners) with the logistic regression (meta-learner).

The meta-model combines their outputs for base models $f_1, f_2, \ldots, f_k$ and is given by

$$\hat{y} = \sigma\left(\sum_{j=1}^{k} w_j f_j(x)\right), \tag{4}$$

where σ is the sigmoid function, and $w_j$is the weight of the j^{th} base model.

The individual base model addressed class imbalance independently using their respective methods, while the meta-model is trained on weighted predictions [20].

C. Cross-Validation Strategy

Train-test splitting can yield unreliable results in imbalanced datasets, as minority classes might not be well-represented in both sets. So, we used stratified cross-validation for model training and testing. Stratified cross-validation ensures that the class distribution in every fold aligns with the overall dataset, resulting in more accurate and more reliable performance metrics [21].

The dataset is divided into k subsets $(D_1, D_2, \ldots, D_k)$ for k-fold cross-validation purposes. The evaluation metric (e.g., F1-score) is computed as

$$Metric_{CV} = \frac{1}{k}\sum_{j=1}^{k} Metric(D_i). \tag{5}$$

The F1 score is the major metric used to evaluate classification models on unbalanced datasets.

$$F1 = 2\frac{Precision.\ Recall}{Precision + Recall}, \tag{6}$$

where precision (P) is the fraction of accurately predicted positive instances out of all instances predicted as positive, which is given by

$$P = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)}, \tag{7}$$

and recall (R) is the fraction of correctly predicted positive instances out of all actual positive instances, which is given by

$$R = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)}. \tag{8}$$

The weighted-average F1-Score is given as

$$F1_{weighted} = \sum_{I=1}^{C} \frac{n_i}{n} F1_i, \tag{9}$$

where n_i is the number of true instances for class i, and n is the total number of instances [22].

In this paper, the 10-fold stratified cross-validation is used to train and evaluate the models across each fold.

D. Bayesian Optimization (for optimizing the Hyperparameter)

Hyperparameter optimization is crucial in enhancing model performance. Since, the traditional grid search and random search methods are computationally expensive and inefficient for large parameter spaces, we utilized Bayesian Optimization (BO) [23]. BO is an efficient technique for optimizing hyperparameters in multiphase flow classification. The objective function $f(\mathbf{x})$ is modelled and optimised using a Gaussian process (GP) and an acquisition function.

BO is used to optimize hyperparameters for all the selected ML models to classify flow regimes. Cross-validation is performed for each set of hyperparameters provided by BO to evaluate the model's performance. The objective function value for those hyperparameters is determined using the cross-validation metric F1-Score.

III. Results and discussion

The flow regime classification is predicted using six ML models: KNN, Naive Bayes, XGBoost, CatBoost, LightGBM, and the Hybrid Stacking model. The selected ML models are trained and tested to classify input features (Superficial gas velocity, superficial liquid velocity, and water cut) into flow regimes. To evaluate the model's performance in classifying flow regimes, the stratified 10-fold cross-validation technique is used. This technique ensured that the class distribution in each fold is preserved, providing a reliable assessment of model performance and reducing risks of overfitting and underfitting of model.

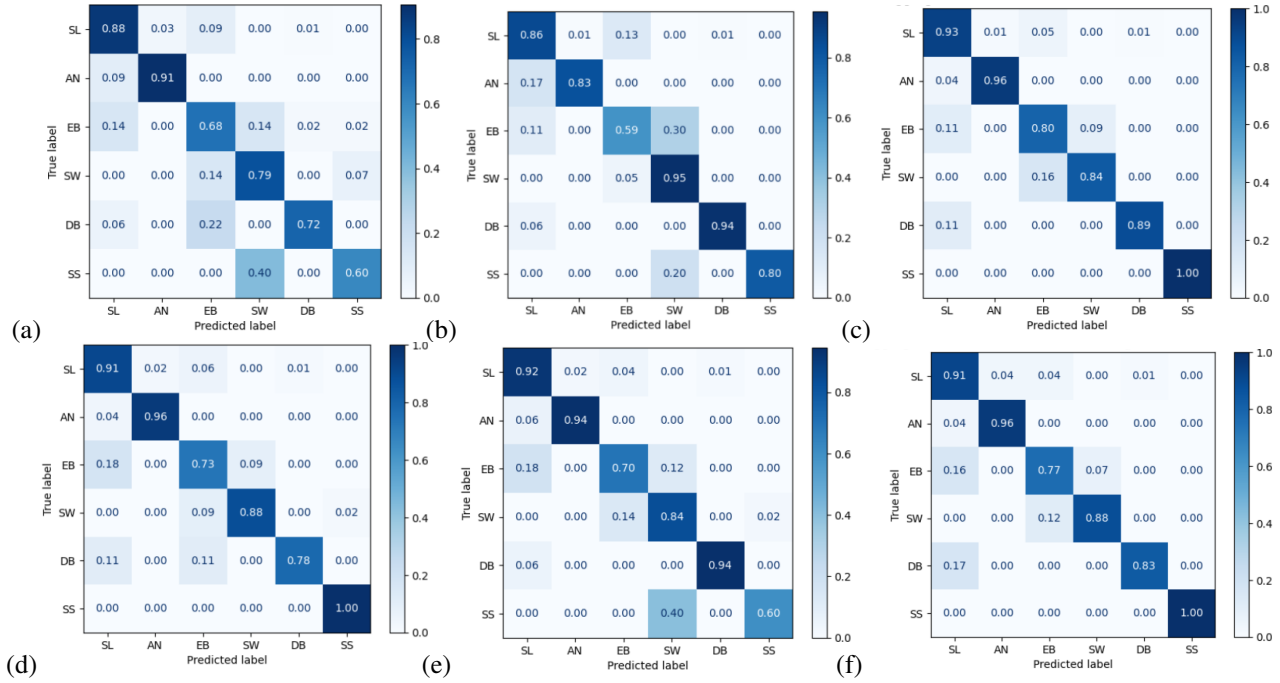

Fig. 3: Confusion matrices for six ML models: (a) KNN, (b) Naive Bayes, (c) CatBoost, (d) LightGBM, (e) XGBoost, and (f) Stacked model, showing their classification performance on flow regimes

TABLE I: Optimized hyperparameters for ML models (LightGBM, CatBoost, XGBoost, KNN, and Naive Bayes)

Algorithm	Hyperparameter	Value
LightGBM	colsample_bytree	1.0
	learning_rate	0.0729
	max_depth	8
	min_child_samples	23
	n_estimators	218
	num_leaves	22
	subsample	0.6
CatBoost	bagging_temperature	1
	border_count	173
	depth	10
	iterations	300
	l2_leaf_reg	1
	learning_rate	0.3
	random_strength	0
XGBoost	colsample_bytree	1.0
	gamma	0
	learning_rate	0.01
	max_depth	10
	n_estimators	300
	reg_alpha	0
	reg_lambda	0.1
	subsample	1.0
KNN	n_neighbors	1
	p	2
	weights	distance
Naive Bayes	var_smoothing	2.18475×10^{-5}

The optimized hyperparameters obtained using the Bayesian optimization are shown in TABLE I. These hyperparameters optimize model performance by balancing the bias-variance tradeoff. LightGBM/XGBoost tuning affects feature selection, learning rate, depth, and regularization for robust learning. CatBoost's bagging, depth, and L2 regularization improve generalization. KNN's distance weighting impacts local decisions, while Naive Bayes' smoothing prevents overfitting in probability estimations. The Bayesian optimization reduced the number of required evaluations by selecting promising hyperparameter values based on prior performance. It also helped in faster convergence and improved accuracy with optimized computational resources. The confusion matrices for the considered ML models, which discusses the model performance across six class labels (SL, AN, EB, SW, DB, and SS) are shown in Fig. 3. In the figure, each individual row represents the true (actual) labels, and each individual column represents the predicted labels. It is clearly depicted that the diagonal elements indicate correctly classified samples, whereas the off-diagonal elements represent misclassifications.

TABLE II: Comparison of ML models' performance metrics

Model	Weighted F1 Score	Avg Training Time per Folds (Sec)	Avg Prediction Time per Folds (Sec)
CatBoost	0.9070	24.5383	0.0013
Stacked Model	0.8936	143.3600	0.0082
LightGBM	0.8866	0.1554	0.0021
XGBoost	0.8760	0.2456	0.0042
KNN	0.8340	0.0012	0.0017
Naive Bayes	0.8264	0.0014	0.0011

The average training and prediction times for the selected ML models are shown in Fig. 4 and Fig. 5, respectively. The comparison of weighted average F1 scores for the ML models is shown in Fig. 6. The classification performance of the machine learning models for multiphase flow regimes is tabulated in TABLE II. The CatBoost model attained the highest weighted average F1 score of 0.9070, while Naive

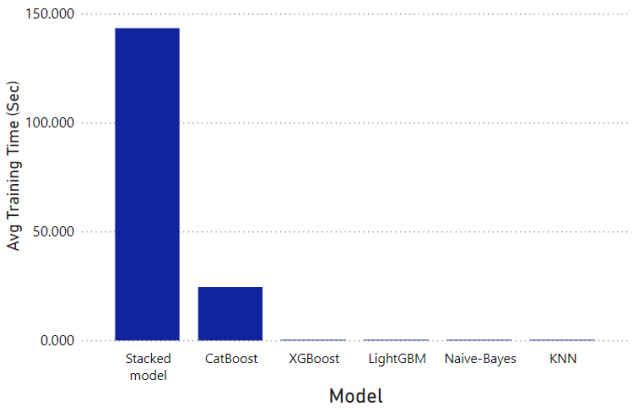

Fig. 4: Comparison of average training time (in seconds) for ML models

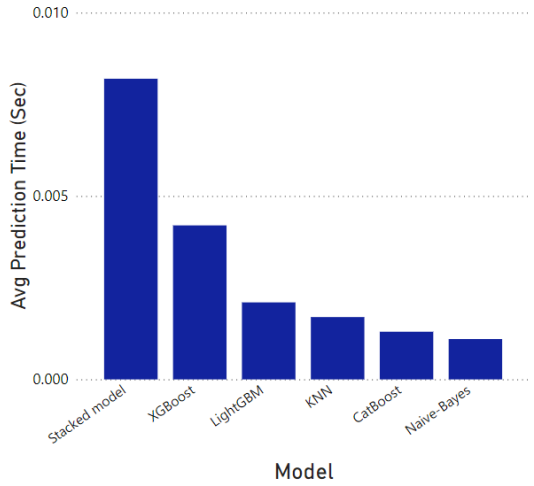

Fig. 5: Comparison of average prediction time (in seconds) for ML models

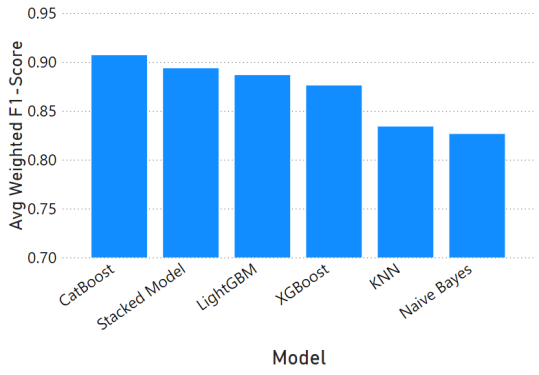

Fig. 6: Comparison of Weighted average F1-Scores for ML models

Bayes performed lowest at 0.8264, illustrating the ensemble learning boosting models' robust performance in handling the class imbalance effectively. The stacked model (combined multiple classifiers) achieved a weighted average F1 score of 0.8936, depicting that ensemble learning enhances classification performance. However, because of complexity, the Stacked Model resulted in a higher average training time (143.36 seconds per fold) and average prediction time (0.0082 seconds per fold) than the other models. LightGBM and XGBoost also performed well, resulting in weighted average F1 scores of 0.8866 and 0.8760, respectively. The gradient-boosting algorithms exhibited a balanced trade-off between accuracy and computational efficiency. LightGBM is reported to be one of the fastest in terms of average training time (0.1554 seconds per fold). It is clearly observed that the class weights are crucial in improving their performance, especially in ensemble models. The inclusion of weights allowed the models to allocate more importance to the minority classes, thereby reducing the bias towards the majority classes and enhancing their ability to classify the imbalanced data. The lowest average training and prediction times are obtained for KNN (0.0012 sec per fold) and Naive Bayes (0.0011 sec per fold), respectively. The results demonstrated that class weights profoundly impacted ensemble models, which significantly improved due to their ability to integrate class weights into their base learners by ensuring equitable learning across all classes. The incorporation of class weights in KNN and NB helped to overcome the inherent limitations in handling imbalanced datasets, resulting in improved classification performance. The accuracy archived by these ML models is higher when compared to deep learning models such as MLP (Average CV F1 score = 0.795) which utilized the same dataset for classification [24]. Incorporating weights helped ML models outperform MLP by adapting to misclassified samples, handling class imbalances, and improving decision-making, ensuring better generalization in multiphase flow classification.

IV. Conclusion

In this paper, the effectiveness of ML models in classifying flow regimes is successfully demonstrated using a stratified 10-fold cross-validation technique. Ensemble learning models such as CatBoost and LightGBM performed well, with weighted average F1 scores of 0.9070 and 0.8866, respectively. It is noticed that the LightGBM model stood out for its excellent balance between accuracy and computational time, with a significantly lower average training time than CatBoost. The XGBoost model also performed well, with lower training and prediction times. These findings highlighted the ability of ensemble methods to capture complex relationships within the data, ensuring robust and reliable classification performance, especially in multi-class imbalanced datasets. KNN and Naive Bayes performed less than ensemble learning methods but proved computationally efficient. Despite their lower F1 scores compared to ensemble methods, the minimal computational cost suits scenarios requiring quick predictions

with limited computational resources. The stacked model achieved a weighted average F1-Score of 0.8936, showcasing the potential of combining multiple algorithms to enhance predictive capabilities at a higher computational cost. High-performing models like CatBoost and Stacked Models required more computation, while faster models like KNN and Naive Bayes sacrifice some accuracy for speed. The LightGBM and XGBoost strike a balance between accuracy and computational efficiency. This study emphasized the importance of accurate flow regime classification, a critical factor in optimizing processes across industries such as petroleum engineering, chemical processing, and multiphase flow simulations. Further, integrating these predictive models into real-time monitoring and control systems can enhance safety, prevent system failures, and minimize downtime. Moreover, the application of machine learning models, particularly those optimized using Bayesian techniques, ensures robust classification even in complex flow conditions. The adaptability of these models allows for seamless integration into industrial automation systems, enabling proactive decision-making and reducing reliance on empirical correlations. Additionally, AI-driven classification enhances energy efficiency, resource utilization, and predictive maintenance, making it invaluable for industries where multiphase flow dynamics play a crucial role.

ACKNOWLEDGMENT

AR was supported by the Department of Chemical Engineering, NIT Andhra Pradesh to carry out the work.

REFERENCES

[1] C. E. Brennen, *Fundamentals of Multiphase Flow*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[2] G. F. Hewitt and N. S. Hall-Taylor, *Annular Two-Phase Flow*. Oxford, U.K.: Pergamon Press, 1970.

[3] Y. Taitel and A. E. Dukler, "A model for predicting flow regime transitions in horizontal and near horizontal gas-liquid flow," *AIChE J.*, vol. 22, no. 1, pp. 47–55, 1976.

[4] G. F. Hewitt, D. N. Roberts, and P. B. Whalley, "Empirical methods for the prediction of flow patterns in two-phase systems," *Int. J. Multiphase Flow*, vol. 5, no. 5, pp. 439–446, 1979.

[5] J. Mandhane, G. Gregory, and K. Aziz, "A flow pattern map for gas-liquid flow in horizontal pipes," *Int. J. Multiphase Flow*, vol. 1, no. 4, pp. 537–553, 1974.

[6] H.-Q. Zhang and C. Sarica, "Unified modeling of gas/oil/water-pipe flow-basic approaches and preliminary validation," *SPE Proj. Facil. Constr.*, vol. 1, no. 2, pp. 1–7, 2006.

[7] N. Petalas and K. Aziz, "A mechanistic model for multiphase flow in pipes," *J. Pet. Sci. Eng.*, vol. 26, no. 3–4, pp. 99–107, 2000.

[8] J. O. Salgado, S. C. Vieira, and A. S. Menezes, "Gamma-ray-based artificial neural network approach for flow regime identification and volume fraction prediction in multiphase flows," *Flow Meas. Instrum.*, vol. 21, no. 3, pp. 328–335, 2010.

[9] V. Nasteski, "An overview of the supervised machine learning methods," *Horizons B*, vol. 4, pp. 51–62, 2017.

[10] M. Al-Naser, M. Elshafei, and A. Al-Sarkhi, "Artificial neural network application for multiphase flow patterns detection: A new approach," *J. Pet. Sci. Eng.*, vol. 145, pp. 548–564, 2016.

[11] D. Wang, H. Guo, Y. Sun, H. Liang, A. Li, and Y. Guo, "Prediction of oil–water two-phase flow patterns based on Bayesian optimization of the XGBoost algorithm," *Processes*, vol. 12, no. 8, p. 1660, 2024.

[12] A. Al-Dogail, R. Gajbhiye, A. AlNajim, and M. Al-Naser, "Dimensionless artificial intelligence-based model for multiphase flow pattern recognition in horizontal pipe," *SPE Prod. Oper.*, vol. 37, no. 2, 2022.

[13] L. M. Al-Hadhrami, S. M. Shaahid, L. O. Tunde, and A. Al-Sarkhi, "Experimental Study on the Flow Regimes and Pressure Gradients of Air-Oil-Water Three-Phase Flow in Horizontal Pipes," *Sci World J*, vol. 2014, no. 1, p. 810527, 2013.

[14] G. King and L. Zeng, "Logistic regression in rare events data," *Political Anal.*, vol. 9, no. 2, pp. 137–163, 2001.

[15] S. Uddin, I. Haque, H. Lu, *et al.*, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Sci. Rep.*, vol. 12, p. 6256, 2022.

[16] I. Rish, "An empirical study of the Naïve Bayes classifier," in *IJCAI 2001 Work. Empir. Methods Artif. Intell.*, 2001.

[17] J. Tanha, Y. Abdi, N. Samadi, *et al.*, "Boosting methods for multi-class imbalanced data classification: An experimental review," *J. Big Data*, vol. 7, p. 70, 2020.

[18] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: An interdisciplinary review," *J. Big Data*, vol. 7, p. 94, 2020.

[19] M. Gan, S. Pan, Y. Chen, *et al.*, "Application of the machine learning LightGBM model to the prediction of the water levels of the Lower Columbia River," *J. Mar. Sci. Eng.*, vol. 9, p. 496, 2021.

[20] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992.

[21] Purushotham, S., & Tripathy, B. K. Evaluation of classifier models using stratified tenfold cross validation techniques. In International conference on computing and communication systems, Berlin, Heidelberg, pp. 680-690, 2011

[22] J. Tanha, Y. Abdi, N. Samadi, *et al.*, "Boosting methods for multi-class imbalanced data classification: An experimental review," *J. Big Data*, vol. 7, p. 70, 2020.

[23] T. T. Joy, S. Rana, S. Gupta, and S. Venkatesh, "Hyperparameter tuning for big data using Bayesian optimization," in *Proc. Int. Conf. Pattern Recognition (ICPR)*, Cancun, Mexico, pp. 2574–2579, 2016.

[24] Alhashem, Mayadah. "Machine Learning Classification Model for Multiphase Flow Regimes in Horizontal Pipes." Paper presented at the International Petroleum Technology Conference, Dhahran, Kingdom of Saudi Arabia, January 2020.