UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

FEATURE ENGINEERING AND HYBRID MODELING IN MACHINE LEARNING: A UNIFIED APPROACH FOR REDUCING COMPUTATIONAL COMPLEXITY IN FLOW PATTERN IDENTIFICATION AND HYDROCARBON PRODUCTION FORECASTING

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

GENE MICHAEL MASK

Norman, Oklahoma

2025

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/341342

FEATURE ENGINEERING AND HYBRID MODELING IN MACHINE LEARNING: A UNIFIED APPROACH FOR REDUCING COMPUTATIONAL COMPLEXITY IN FLOW PATTERN IDENTIFICATION AND HYDROCARBON PRODUCTION FORECASTING

A DISSERTATION APPROVED FOR THE

MEWBOURNE SCHOOL OF PETROLEUM AND GEOLOGICAL ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Xingru Wu, Chair

Dr. Mashhad Fahes

Dr. Hamidreza Karami Mirazizi

Dr. Rouzbeh Ghanbarnezhad-Moghanloo

Dr. Charles D. Nicholson

© Copyright by GENE MICHAEL MASK 2025

All Rights Reserved.

Acknowledgements

I am deeply grateful to my advisor, Dr. Xingru Wu, for his unwavering support, guidance, and wisdom throughout every stage of this dissertation. His expertise and encouragement have been invaluable in shaping this research and my life.

I extend my heartfelt thanks to my esteemed committee members: Dr. Mashhad Fahes, Dr. Hamidreza Karami Mirazizi, Dr. Rouzbeh Ghanbarnezhad-Moghanloo, and Dr. Charles D. Nicholson. Their confidence in me, insightful questions, and willingness to share their expertise have significantly enriched this work.

I am indebted to my family—my wife, Wanwaree, and my children, Talia, Garron, and Banyan Mask—for their unwavering support and understanding during this challenging journey. Their love and encouragement have been my pillars of strength.

I am grateful to the University of Oklahoma for providing me with a nurturing environment and continuous support, making it feel like a home away from home. The resources and opportunities offered by the university have been instrumental in my academic endeavors.

To all those who have contributed to this journey, whether directly or indirectly, your support has been invaluable, and I am deeply grateful for it.

Thank you.

Table of Contents

Acknowledgments iv

List of Tables ..... ix-xi

List of Figures ..... xii-xviii

Abstract ..... xix-xxi

Chapter 1: Introduction 1

1.1 Problem Statement 1-2

1.2 Research Objective 2-4

1.3 Organization of Dissertation 4-5

Chapter 2: Literature Review 6

2.1 Fundamentals and Applications 6-7

2.1.1 Dimensional Analysis 7-16

2.1.2 Transfer Learning 16-19

2.1.3 Fluid Flow Pattern Recognition 19-22

2.1.4 Hydrocarbon Production Forecasting 23-35

2.1.5 Machine Learning 35-46

2.1.6 Data Preprocessing and Validation 46-53

2.1.7 Model Evaluation Metrics 53-55

Chapter 3: Machine Learning Two-Phase Flow Pattern Identification 56

3.1 Research Motivation 56-57

3.2 Machine Learning Model Details 57-58

3.2.1 Experimental Data 58-61

3.2.2 Dimensional Reduction 61-63

3.2.3 Machine Learning Gas-Liquid Flow Pattern Prediction Model ... 63-64

3.3 Experimental Procedure 64

3.3.1 Transforming to Dimensionless Predictors 65

3.3.2 Feature Extraction 66-70

3.3.3 Experimental Model Design 70-71

3.4 Results and Discussion 71-73

Chapter 4: Machine Learning Enhanced Hydrocarbon Production Forecasting 74

4.1 Research Motivation 74-75

4.2 Methodology 75-79

4.3 Experimental Data 79-81

4.3.1 Data Preprocessing 81-86

4.3.2 Dimensional Analysis for Feature Reduction 86-88

4.4 Transfer Learning – Scalar to Vector Machine Learning Model 88-89

4.4.1 Static Machine Learning Model Tuning and Validation 90

4.4.2 Static EUR Machine Learning Model 90-91

4.4.3 User Transfer Learning 91-94

4.5 Vector Machine Learning Model 94-95

4.5.1 Machine Learning Cumulative Production Forecasting 95-98

4.5.2 Validate Transfer Learning – Vector Model 99-101

4.6 Implementing Machine Learning Assisted – Decline Curve Analysis .. 101-103

Chapter 5: Conclusions and Future Work 104

5.1 Conclusions..... 104-105

5.2 Future Work 105-106

Acknowledgement 107

Nomenclature 107

References 108-116

Appendix A. The principle of absolute significance of relative magnitude..... 117

Appendix B. Buckingham Π Theorem 118-119

Appendix C. Dimentionless Decline Curve Analysis Equations 120-122

Appendix D. Empirical Hydrocarbon Production Comparison..... 123

Appendix E. Machine Learning Expressions 124-128

Appendix F. Gadiant Boosting Algorithm..... 129

Appendix G. SVR Mathematical Formulation 130-131

Appendix H. Scaling Metrics 132-134

Appendix I. Statistical Evaluation Metrics..... 135-138

Appendix J. Mathematical Derivation of Fluid Flow Dimensionless Groups..... 139-141

Appendix K. Data Analysis Flow Patterns 142-146

Appendix L. Reservoir Property Calculations 147-149

Appendix M. Dimensional Analysis of Reservoir Properties 150-153

Appendix N. Test Train Results for EUR ML Model Integrated with DA..... 154-155

Appendix O. Scalar (EUR) Regression Model Hyperparameters..... 156-157

Appendix P. Temporal Feature Vector 158

Appendix Q. Typer Curve and MAE Bounds 159-160

Appendix R. Machine Learning Automated Dimensional Analysis 161-163

List of Tables

Table 1. Fundamental quantities form the basis of physical measurement, each defined by a primary dimension and an assigned unit. These quantities remain consistent across measurement systems, ensuring standardization in scientific analysis 8

Table 2. Derived quantities obtained from fundamental quantities through mathematical relationships, enabling a deeper understanding of physical phenomena. The derived quantities essential for scientific measurement and technological advancements..... 9

Table 3. Buckingham Π Theorem enables a systematic approach to forming dimensionless groups for scaling and similarity analysis. The steps ensure dimensional consistency in physical equations (Buckingham, 1914)..... 10

Table 4. Inspectional Analysis systematically reduces governing equations into dimensionless form to simplify scaling and similarity analysis. The steps to apply IA to derive meaningful dimensionless groups is presented (Ruark, 1935)..... 12

Table 5. The evolution of two-phase flow modeling is summarized, highlighting the transition from empirical flow regime maps to AI-driven predictive models (Baker, 1954; Beggs and Brill, 1973; Gould et al., 1974; Mandhane et al., 1974; Taitel and Dukler, 1976; Barnea et al., 1982a; 1982b; Li et al., 2014; Brito et al., 2015; Al-Safran and Al-Qenae, 2018; Mask et al., 2019) 21

Table 6. The experimental studies on gas-liquid two-phase flow in pipes, details the fluid properties, pipe geometries, pressure conditions, and institutions that contributed to the dataset. These studies provide a broad range of flow conditions for developing and validating ML-based flow pattern prediction models (Pereya et al., 2011). 60

Table 7. Dimensionless Numbers: Definitions and Physical Interpretations 62

Table 8. Confusion matrix for the XGB tree model using 5-fold, 10-repeated CV showing classification performance. The model achieves $\geq 90\%$ accuracy for most patterns, with intermittent flow having the highest accuracy and bubble flow the lowest 71

Table 9. Performance metrics of tree-based models using 5-fold and 10-fold, 10-repeated cross-validation. All models demonstrated over 90% accuracy using only four dimensionless predictors..... 72

Table 10. The statistical values for hydraulic fracturing treatment parameters used in the dataset help define the distribution of well characteristics used for training the machine learning model, ensuring that production forecasts align with real-world completion conditions 80

Table 11. The statistical overview of key reservoir characteristics in the dataset highlights the variability in reservoir properties. Understanding these distributions is essential for modeling production behavior and optimizing machine learning prediction 81

Table 12. Dimensionless parameters derived using dimensional analysis combine and simplify the predictive features 88

Table 13. R^2 test and train scores for Repeated K-Fold (5-fold, 10-repeated) CV in predicting EUR indicate that the DA model outperforms the PCA model as a dimensionality reduction technique 91

Table D1. The comparative analysis of empirical DCA models, highlighting their strengths and weaknesses. Understanding these trade-offs is essential for selecting the most suitable model for unconventional reservoirs (Yehia et al., 2022) 123

Table F1. Generalized boosting of decision trees (Friedman, 2002) 129

Table O1. Final static ML model hyperparameters (Pedregosa et al., 2011)..... 157

Table R1. The possible combinations of completion related Π groups formulated using the Buckingham Π theorem. These groups help identify dominant physical interactions governing hydraulic fracturing performance, enabling more efficient and scalable production forecasting models 162

Table R2. The possible combinations of reservoir related Π groups. These dimensionless groups facilitate the development of generalized machine learning models, ensuring applicability across diverse reservoir conditions while preserving underlying flow physics 163

List of Figures

Figure 1. Schematic of flow regime phenomena during shale gas production. Point A depicts fracture linear flow, B depicts bilinear flow, C depicts formation linear flow, D depicts transition from linear flow to radial flow, E depicts pseudo-radial flow (Childers and Wu, 2020) 24

Figure 2. The variability in production forecasts generated by different decline curve models when trained on only 20% of the available production data. Significant deviations between models highlight the sensitivity of forecasting methods to data availability and model assumptions. The study emphasizes the importance of selecting an appropriate DCA model based on data volume, outlier treatment, and the model's inherent assumptions Yehia et al. (2022)..... 32

Figure 3. The DCA forecasts improve when fitted with 80% of the production data. As a result, many of the DCA models begin to converge more closely to the actual production trend (blue dots), reducing variability in long-term predictions. This highlights the importance of data availability in enhancing the reliability of production forecasts, as shown in Yehia et al. (2022)..... 33

Figure 4. Transition boundaries and observed flow patterns in gas-liquid two-phase flow, including dispersed bubble, stratified smooth, stratified wavy, annular, intermittent, and bubble. These boundaries illustrate the classification of flow regimes based on experimental data (Pereya et al., 2011) 59

Figure 5. The even distribution of mFr_l after Box-Cox transformation, indicates its suitability as a predictor, while distinct flow pattern groupings highlight its role in displaying flow regime transition 67

Figure 6. The slight right skew of mFr_g after Box-Cox transformation contrasts with the more uniform distribution of mFr_l, indicating differences in flow regime behavior and supporting its role as a predictive feature 67

Figure 7. Distinct clustering of flow patterns gas-liquid two-phase flow for air and water in a horizontal pipe indicates transition boundaries, supporting the use of mFr_l vs. mFr_g as key features for flow regime classification 68

Figure 8. The transition boundaries of the mFr_l vs. mFr_g for air and water in all pipes between flow patterns become less distinct, indicating increased complexity when incorporating additional flow orientations 68

Figure 9. The transition boundaries of the mFr_l vs. mFr_g of all fluids plot shows flow patterns become less distinguishable, indicating that additional complexities arise when generalizing across different flow conditions and fluid types 69

Figure 10. Distributions of key completion parameters illustrating how some variables present skewed distributions while others appear more normally distributed. Higher-productivity wells cluster in distinct parameter ranges, highlighting potential “sweet spots” and offering insights for optimizing completion designs 83

Figure 11. Histograms of key treatment parameters revealing how different operating ranges correlate with production outcomes. These distributions show that certain

properties are more often associated with higher productivity, offering insights that can guide well design optimization 84

Figure 12. Distribution of reservoir properties highlights their impact on production performance. The trends indicate that higher-producing wells are associated with specific ranges of these key reservoir characteristics. The permeability distribution exhibits a bimodal pattern, reflecting contributions from two distinct reservoirs with differing permeability characteristics..... 85

Figure 13. The histogram highlight relationship between key reservoir and fluid properties with production performance. The bimodal distribution in fluid density and viscosity reflects differences between oil and gas formations. The latitude and longitude distributions indicate that the analyzed wells are in distinct geological regions further emphasizing formation-dependent variations in production characteristics 86

Figure 14. Presents a parity plot comparing predicted and observed EUR values, where the red line represents the ideal 1:1 correlation and the black dots denote model predictions. Deviations from the red line highlight discrepancies, offering insights into model accuracy, bias, and potential sources of error 93

Figure 15. The feature importance ranking for production forecasting show the top contributing predictive features Π_6, longitude, and latitude, indicating the significant influence of spatial location and select dimensionless groups on production 94

Figure 16. Cumulative oil production forecast for a Wolfcamp well using a machine learning model trained on both Montney and Wolfcamp data. The ML production curve (green dots) closely follows the operator's production curve (black triangles), but exhibits

a slight underprediction, particularly in later months. This suggests that the inclusion of Montney data introduces discrepancies in production behavior, due to geological and completion differences between the formations 97

Figure 17. Cumulative oil production forecast for a Wolfcamp well using a ML model trained exclusively on Wolfcamp data. The ML production curve (green dots) aligns closely with the operator's production curve (black triangles), demonstrating improved accuracy compared to the mixed-dataset model. This suggests that training on formation-specific data enhances the model's ability to capture unique reservoir characteristics and production 98

Figure 18. The impact Π_6 on production trends reinforces its significance in improving model forecasts, validating its role as the most influential feature in transfer learning. This result aligns with the feature importance ranking in **Figure 15**, confirming Π_6 as a key variable in shaping production predictions and validating knowledge transfer from scalar to vector ML model 100

Figure 19. Sensitivity analysis of cumulative oil production per 1000 ft, demonstrating the minimal impact of predictor $\Pi_3$on production forecasting. Variations in Π_3 result in negligible differences in the predicted production curves, suggesting that the model could be further simplified by removing Π_3 without compromising accuracy 101

Figure 20. Comparison of cumulative oil production forecasts illustrating the MLA-DCA approach. Here, the ML model (green dots) closely matches the operator’s production curve (black triangles) for the initial production phase, while the empirical component (purple stars) extends the forecast into later months. This highlights the power of

integrating machine learning and traditional decline curve analysis to generate accurate, long-term production forecasts..... 103

Figure K1. The flow pattern distribution for the $Eo^{0.25} * \cos \measuredangle_{pipe}$ highlights the distribution of flow patterns across different values of the Eo, capturing the influence of buoyancy and inclination on flow regime transitions. The clustering of specific flow patterns at distinct values supports its use as a predictive feature in flow modeling 142

Figure K2. The flow pattern distribution for $Eo^{0.15} * \sin \measuredangle_{pipe}$ highlights the influence of gravitational and interfacial forces on flow regimes, showing distinct clustering of flow patterns at specific values. The distribution supports the relevance of this dimensionless parameter in predicting multiphase flow behavior 143

Figure K3. Plot of Π_2 and Π_1 with existing flow patterns in gas-liquid two-phase flow for all fluids in all pipes. Unlike traditional flow regime maps that show continuous phase transitions, the discrete clustering of flow patterns here suggests that these dimensionless groups act as strong separators of multiphase flow behavior. However, the lack of smooth transitions between patterns indicates that additional considerations may be needed to fully capture the complexity of real-world flow regimes..... 144

Figure K4. Plot of Π_3 and Π_1 with existing flow patterns in gas-liquid two-phase flow for all fluids in all pipes. The structured clustering highlights the ability of these dimensionless groups to differentiate multiphase flow behavior. However, the qualitative nature of the flow regime transitions may require further validation to ensure that gradual phase shifts are adequately represented 145

Figure K5. Plot of Π_3 and Π_2 with existing flow patterns in gas-liquid two-phase flow for all fluids in all pipes. The structured and linear clustering of flow patterns suggests a strong correlation between these dimensionless parameters, indicating their predictive capability in distinguishing flow regimes. However, unlike traditional flow regime maps, this visualization lacks smooth transitions between patterns, which may require additional validation or alternative visualization techniques to better represent gradual phase changes 146

Figure N1. The performance of various machine learning models in predicting EUR, with tree-based models (RF, ET, and XGB) demonstrating consistent and strong predictive test performance (R^2>0.6) The relatively higher and stable scores of these models indicate their robustness in capturing underlying data patterns. Other models, including neural networks (MLP, RNN, CNN) and linear regressors, exhibit greater variability, highlighting differences in model generalization. The results suggest that tree-based models provide reliable predictions, making them well-suited for practical applications in hydrocarbon production forecasting 154

Figure N2. The training performance of various machine learning models, highlighting the exceptional generalization of tree-based models (RF, ET, and GBR), which achieve consistently high training scores (R^2>0.8). The high scores indicate that these models effectively capture underlying data patterns with minimal overfitting. In contrast, linear models and some neural network-based models (MLP, RNN, CNN) exhibit greater variability, suggesting differences in model complexity and adaptability to training data. The results reinforce the effectiveness of tree-based models for predictive accuracy in hydrocarbon production forecasting. 155

Figure Q1. The Mean Absolute Error bounds provide an uncertainty range, ensuring that more than 80% of the predicted forecasts fall within the ± MAE margin, reinforcing the model's reliability..... 160

Figure Q2. The Mean Absolute Error provides an uncertainty range, reinforcing the need for additional historical data to enhance long-term predictive accuracy and reduce forecast variability 160

Abstract

Engineering applications require predictive models that not only capture complex, nonlinear relationships but also remain physically consistent and generalizable across varying operational conditions. Achieving this balance is challenging, as existing models each have inherent limitations. Many deterministic models, including empirical and mechanistic approaches, assume fixed conditions and struggle to adapt to real-world complexity. Probabilistic models introduce computational challenges and often lack direct ties to physical laws. Machine learning (ML) approaches also face challenges in petroleum engineering due to limited physical interpretability and difficulty generalizing across diverse geological and operational conditions.

Empirical models, often based on curve fitting or historical data, are fast but tend to assume homogeneous reservoirs and fixed fluid properties. This makes them unreliable under changing reservoir and operational conditions. Mechanistic models, grounded in first-principles physics, improve interpretability but typically require extensive calibration and high-fidelity input data, making them less practical for real-time applications. Both types of models struggle with complex flow behavior, particularly in unconventional plays.

Accurately predicting pressure profiles along wells is essential for production optimization and well integrity analysis. The calculation is predicated on the accurate classification of flow patterns. However, empirical and mechanistic models often fail to capture complex flow patterns, such as slug, churn, and annular flow, that cause significant pressure fluctuations. These flow regimes introduce operational inefficiencies and risks that are difficult to model without sacrificing accuracy or computational feasibility.

Probabilistic models, such as Monte Carlo simulations, Bayesian inference, and stochastic modeling, address uncertainty by generating distributions of possible outcomes rather than a single forecast. While useful for risk assessment, they often lack physical constraints, and their accuracy depends on well-defined input distributions and high-quality data. In applications like probabilistic reserve estimation, geological heterogeneity and parameter uncertainty make it difficult to validate input assumptions. As a result, probabilistic models alone are typically insufficient for robust forecasting and are often used in combination with empirical or physics-based methods.

To overcome these limitations, this study proposes a unified machine learning framework that integrates feature engineering, dimensional analysis (DA), transfer learning (TL), and decline curve analysis (DCA). This hybrid approach enhances both flow pattern classification and hydrocarbon production ML forecasting while maintaining computational efficiency and physical consistency.

Conventional mechanistic models often achieve less than 85% accuracy in predicting pressure profiles across different flow regimes. The study applied DA to derive three dimensionless predictors using the Buckingham Π Theorem. Embedding these predictors into the ML model ensures consistency with underlying physics and improves generalization. Among the evaluated models, the Extreme Gradient Boosted Random Forest (XGB-RF) achieved the highest accuracy of 93% (Kappa = 0.90), outperforming mechanistic models. This highlights the effectiveness of physics-guided feature engineering in refining ML predictions for two-phase flow pattern classification.

For production forecasting in multi-fractured horizontal wells (MFHWs), traditional rate-based models often fail due to early-time data sparsity and operational variability. The

Machine Learning Assisted – Decline Curve Analysis (MLA–DCA) framework addresses this by integrating DA, TL, and DCA. DA simplifies input complexity while preserving physical meaning. TL enables knowledge transfer from a pre-trained scalar model to a more adaptive vector-based ML model, optimizing both algorithm and hyperparameter selection. DCA complements the ML model for late-time forecasts, where data becomes increasingly sparse.

The MLA–DCA framework achieves $R^2 \geq 0.80$ in cumulative production forecasting, demonstrating its effectiveness in balancing accuracy, generalizability, and efficiency. By minimizing feature space, training requirements, and computation time, the framework is scalable for real-world deployment. Feature engineering ensures physical alignment, while hybrid modeling improves performance across diverse flow regimes, reservoir types, and operational conditions.

This study presents a physics-informed, data-driven framework that unifies machine learning with engineering fundamentals. The proposed hybrid approach enhances flow pattern identification and production forecasting, advancing the role of ML in petroleum engineering decision-making. By ensuring that models are accurate, interpretable, and efficient, this work provides a robust solution for tackling complex challenges in petroleum engineering applications.

Chapter 1: Introduction

1.1 Problem Statement

Modern petroleum systems, particularly in unconventional reservoirs and multi-fractured horizontal wells, present a high degree of geological heterogeneity, nonlinear flow behavior, and rapidly changing operational conditions. These complexities demand predictive models that are not only accurate and efficient but also physically interpretable and adaptable. However, most existing approaches fall short when applied to these dynamic environments.

Traditional deterministic models, including both empirical and mechanistic types, rely on idealized assumptions that break down under real-world variability. They are often calibrated for specific operating envelopes and fail to scale across different reservoir types, well configurations, or flow regimes. Although numerical simulations offer greater physical fidelity, their reliance on finely tuned input parameters and intensive computation limits their practical use, particularly in time-sensitive applications.

Probabilistic models offer a statistical approach to uncertainty but introduce their own challenges, most notably, a dependence on accurate input distributions and a frequent detachment from physical laws. This can result in outputs that, while probabilistically valid, are not always consistent with engineering principles, limiting their credibility in high-stakes decisions such as reserve estimation and production forecasting.

Meanwhile, machine learning has shown promise in handling nonlinear and high-dimensional data. Yet, most ML models operate as black boxes, lacking transparency, physical grounding, or the ability to generalize beyond their training domain. This creates

a disconnect between predictive accuracy and engineering trustworthiness, an issue especially acute in safety-critical applications like flow pattern classification and well performance forecasting.

There is a pressing need for a unified ML modeling framework that integrates the strengths of ML with physical interpretability, scalability, and computational efficiency. Such a framework should not only improve forecasting accuracy but also enhance trust and usability in real-world petroleum engineering workflows, especially where data is sparse, conditions are variable, and interpretability is essential.

1.2 Research Objective

The aim of this research is to develop a unified, physics-informed machine learning framework that enhances the accuracy, interpretability, and scalability of predictive ML models used in petroleum engineering. This study specifically targets two critical challenges in the field: the accurate classification of two-phase flow patterns and reliable forecasting of hydrocarbon production.

The study seeks to achieve several interrelated and measurable objectives that will guide the design of the methodology, data collection, and analysis. First, the research aims to improve flow pattern classification by applying machine learning models trained on dimensionless features derived from dimensional analysis. The objective is to create a classification model that not only achieves a higher accuracy than traditional empirical or mechanistic models, surpassing the conventional threshold of 85%, but also remains generalizable across diverse operating conditions, fluid systems, and conduit geometries.

The second objective is to enhance hydrocarbon production forecasting through the integration of DA, TL, and DCA into a cohesive framework called Machine Learning Assisted – Decline Curve Analysis. The goal is to reduce input complexity, optimize training efficiency, and increase adaptability by utilizing scalar-based models to inform hyperparameter selection in vector-based production forecasts. This approach is designed to address the limitations of existing ML models, which often require large datasets and struggle with overfitting or poor late-time predictions. By incorporating empirical models into the forecasting pipeline, the study further aims to extend predictive reliability into periods of data scarcity, ensuring robustness during late-stage production decline.

To achieve these objectives, the study will investigate the following research questions: How can dimensional analysis enhance the physical consistency and generalization of ML models for flow pattern recognition? Can transfer learning be used to minimize the data and computation requirements of production forecasting models without sacrificing accuracy? What is the optimal strategy for integrating DCA models with machine learning in a way that complements each method's strengths while mitigating their respective weaknesses?

These objectives and guiding questions will shape the structure of the research, from feature engineering and model development to validation and performance evaluation. They provide a focused and coherent roadmap for the study, ensuring that each phase, data collection, model training, and interpretation, contributes meaningfully to solving the core challenges identified. This research aims to deliver a practical, scalable, and

scientifically grounded solution that supports petroleum engineers in making informed, data-driven decisions under complex and uncertain conditions.

1.3 Organization of Dissertation

This dissertation is structured into five chapters, each systematically built upon fundamental concepts and methodologies to advance machine learning applications in petroleum engineering applications.

Chapter 1 establishes the research problem statement, research objective, and order of the dissertation. It provides a high-level overview of the study and its highlights integrating machine learning with dimensional analysis, transfer learning, and empirical models.

The literature review in Chapter 2 presents a comprehensive examination of the fundamental concepts underpinning this research. It explores key topics such as dimensional analysis, transfer learning, fluid flow pattern recognition, hydrocarbon production forecasting, and machine learning techniques. This chapter contextualizes existing methodologies, outlining their significance in petroleum engineering applications. It also discusses dataset preprocessing and statistical evaluation metrics to validate ML models in petroleum engineering applications.

Chapter 3 introduces a two-phase fluid flow pattern identification using ML models. This chapter verifies that dimensionless features, when used as input to classification models, can effectively predict two-phase fluid flow patterns. The study demonstrates that integrating dimensional analysis into machine learning reduces the input matrix size and model complexity while maintaining high predictive accuracy. The chapter details the

experimental procedure, transformation to dimensionless predictors, feature extraction, and model design, followed by an in-depth discussion of results.

Chapter 4 focuses on the development of ML models for hydrocarbon production forecasting. This chapter provides a step-by-step methodology for constructing and validating machine learning models, enabling replication with new datasets, and subject matter expertise. It includes data preprocessing techniques, dimensional analysis for feature reduction, and model development strategies. Novel transfer learning methods were extracted from scalar (EUR) model to optimize vector (cumulative production over time) model. The chapter systematically evaluates EUR models through hypothesis testing, comparative analyses, and algorithm performance assessments. Then followed with an in-depth validation of transfer learning for cumulative production forecasting and an assessment of model reliability. The chapter concludes with the implementation of Machine Learning Assisted – Decline Curve Analysis showcasing its role in extending ML predictions for late-time forecasting.

In Chapter 5, conclusions and future work present the key research findings and their implications. It highlights the contributions of integrating DA, TL, and empirical models into machine learning for two-phase flow pattern identification and hydrocarbon production forecasting. The chapter also outlines potential avenues for future research, encouraging further advancements in ML applications for petroleum engineering.

This structured approach ensures a logical flow from fundamental concepts to advanced methodologies, providing a clear framework for the study’s contributions to machine learning applications in petroleum engineering.

Chapter 2: Literature Review

2.1 Fundamentals and Applications

A comprehensive review of the foundational principles of dimensional analysis, transfer learning, fluid flow pattern recognition, hydrocarbon production forecasting, and machine learning serves as the basis for this study. These interdisciplinary domains, deeply rooted in mathematics, physics, and computer science, provide critical frameworks for advancing both theoretical understanding and practical innovation. By examining the methodologies and theoretical underpinnings of these fields, this review establishes the necessary context for integrating them into modern engineering solutions. Their enduring relevance is evident across a broad range of scientific and engineering disciplines, where they continue to drive progress and enhance analytical rigor.

Within the evolving landscape of petroleum engineering, the integration of such advanced techniques has become essential for efficient exploration, production optimization, and reservoir management. Machine learning empowers engineers to derive actionable insights from complex, high-dimensional datasets, supporting tasks such as predictive modeling, anomaly detection, and real-time operational decision-making. Dimensional analysis provides a systematic means of deriving physically consistent models and translating laboratory-scale experiments to field-scale applications, thereby improving the generalizability of predictions. Transfer learning enables the application of pre-trained models across domains or datasets, significantly reducing the need for extensive labeled data, a frequent limitation in petroleum engineering contexts. This literature review therefore explores not only the theoretical foundations of these methods but also their specific applications within petroleum engineering, drawing on recent research

developments and case studies to illustrate their practical impact and potential for future innovation.

2.1.1 Dimensional Analysis

Dimensional analysis is utilized to identify relationships and scaling criteria between physical quantities and the phenomena studied. DA generates nondimensional parameters (dimensionless groups) to assist in the design of experiment (physical and/or numerical), report experimental results, obtain scaling laws to predict prototype performance from model performance, and predict trends in the relationship between parameters (White, 2019). DA is a method used to deduce information about a phenomenon, assuming that given certain variables, the phenomenon can be represented by a dimensionally correct equation (Langhaar, 1951). DA reduces a complex physical problem to the simplest form prior to obtaining a quantitative answer (Sonin, 2001). The dimensionless groups obtained from DA represent the governing equations of the phenomenon which may be unknown or difficult to work with. The derived dimensionless groups reduce the number of variables to inspect the phenomenon. In its simplest form, DA checks that a formula makes sense (Attenborough, 2015). The three techniques of dimensional analysis, appropriate scaling, and intelligent estimation of the order of magnitude very often provide nine-tenths of what one can hope to ever know about a problem without actually solving it; the remaining one-tenth requires painstaking mathematics and or lots of computer time (Astarita, 1997).

A dimension is a measure by which a physical variable is conveyed quantitatively, and a unit is a way of assigning a number to that quantitative dimension (White, 2019). Primary (fundamental) dimensions represent the type of base (physical) quantity for purposes of

classification or differentiation. Fundamental quantities seen in **Table 1** are the foundation of physical measurement.

Table 1. Fundamental quantities form the basis of physical measurement, each defined by a primary dimension and an assigned unit. These quantities remain consistent across measurement systems, ensuring standardization in scientific analysis.

Physical Quantity	Dimension	SI unit	SI Symbol
mass	M	kilogram	kg
length	L	meter	m
time	T	second	s
temperature	Θ	kelvin	K
electric current	I	ampere	A
luminous intensity	C	candela	cd
amount of substance	N	mole	mol

Bridgman (1922) explained that the principle of absolute significance of relative magnitude in **Appendix A**, a defining attribute of all systems of measurement, is the ratio of all physical quantities being constant and independent of the units used to measure them. Secondary (derived) quantities obtained from mathematical formulas are composed of numerical values expressing physical quantities of the first kind. **Table 2** shows some commonly derived quantities. These quantities provide a better understanding of the physical phenomena by combining fundamental quantities. It is paramount to understand and measure these derived quantities to advance scientific technology.

Table 2. Derived quantities obtained from fundamental quantities through mathematical relationships, enabling a deeper understanding of physical phenomena. The derived quantities essential for scientific measurement and technological advancements.

	Physical Quantity	Dimensions	SI Symbol
Geometry	Area	L^2	m^2
	Volume	L^3	m^3
	Second moment of area	L^4	m^4
Kinematics	Velocity	LT^{-1}	$m \cdot s^{-1}$
	Acceleration	LT^{-2}	$m \cdot s^{-2}$
	Angular velocity	T^{-1}	s^{-1}
	Flow Rate	L^3T^{-1}	$m^3 \cdot s^{-1}$
	Mass flow rate	MT^{-1}	$kg \cdot s^{-1}$
Dynamics	Force	MLT^{-2}	$kg \cdot m \cdot s^{-2}$
	Torque, moment	ML^2T^{-2}	$kg \cdot m^2 \cdot s^{-2}$
	Work, energy	ML^2T^{-2}	$kg \cdot m^2 \cdot s^{-2}$
	Power	ML^2T^{-3}	$kg \cdot m^2 \cdot s^{-3}$
	Stress, pressure	$ML^{-1}T^{-2}$	$kg \cdot m^{-1} \cdot s^{-2}$

Sonin (2001) stated that at the heart of DA is the concept of similarity. Similarity refers to some equivalence between two objects (e.g., model and prototype) or different phenomena. Mathematically, similarity is a transformation of variables which reduces the number of independent variables to a problem (Sonin, 2001). For scaling to occur, similarity conditions must be met for complete similarity. Geometric (scaling of shape), kinematic (scaling of velocity), and dynamic (scaling of force and torque) similarity must be observed. Dimensional independence is maintained when scaling variables are purely geometric, kinematic, and dynamic quantities (Buckingham, 1914; Bridgman, 1922; Langhaar, 1951). Dimensional analysis allows for the derivation of dimensionless groups

to maintain dimensional similarity. Buckingham Π Theorem and Inspectional Analysis (IA) are two methods widely used to implement dimensional analysis.

Buckingham's (1914) work *On physically similar systems, illustrations of the use of dimensional equations* is the basis for Buckingham Π Theorem in **Appendix B**, which documents one of the first well-known methods for utilizing dimensional analysis. The principle of dimensionless homogeneity states that all terms of a physical expression must have the same dimensions for an equation to be dimensionally correct. Therefore, dimensional representations must be multiplicative based on the principle of absolute significance of relative magnitude and the product theorem. **Table 3.** summarizes the steps to Buckingham Π Theorem.

Table 3. Buckingham Π Theorem enables a systematic approach to forming dimensionless groups for scaling and similarity analysis. The steps ensure dimensional consistency in physical equations (Buckingham, 1914).

Step 1	Identify a complete relevant set of n independent variables.
Step 2	List the set of k fundamental dimensions of n variables.
Step 3	Determine the number of dimensionless groups: $i = n - k$.
Step 4	Select k dimensionally independent scaling variables.
Step 5	Employ algebraic theory to create dimensionless groups.
Step 6	Examine dimensionless groups for meaningful equations.

Following Buckingham's work, Ruark (1935) proposed using variables of the physical conditions, boundary conditions, and or initial conditions of a problem by transforming the variables into dimensionless groups. Simple inspection then shows how the dimensionless groups are related using this method Ruark (1935) named inspectional

analysis. Inspectional analysis combines all governing differential equations that describe the process into a single (general) differential equation to create dimensionless scaling groups that have physical meaning and are scalable (Loomis and Crowell, 1964). The process yields a complete and possibly reduced number of dimensionless groups. Geertsma et al. states (1956) that the resulting dimensionless groups are categorized into independent groups (e.g., dimensionless length, time, etc.), dependent groups (measured variables), and similarity groups (independent constant groups).

Given that all physical equations are relations between independent variables, dependent variables, and constants, the total number of variables minus the number of equations determines the number of independent variables, which can be converted to independent groups by dividing them by some value of the same dimension, characteristic of the system (Geertsma et al., 1956). The dependent variables and constants are divided by one or more equations by a selected constant to create the dimensionless dependent and dimensionless similarity groups. The boundary and initial conditions are handled similarly, assuming the equations are differential equations (Geertsma et al., 1956). The IA process yields a complete set of dimensionless groups that physically describe the problem, but the method becomes rigorous and convoluted when dealing with many complex differential equations. The steps listed in **Table 4** can be followed to identify the complete set of dimensionless groups using the IA method. As Ruark (1935) noted in using IA, the operator only needs to perceive that the dimensionless variables and combinations of dimensional constants which occur in the integrated (general) form of the equations are exactly those occurring in the original (governing) differential equations, boundary conditions, and or initial conditions.

Table 4. Inspectional Analysis systematically reduces governing equations into dimensionless form to simplify scaling and similarity analysis. The steps to apply IA to derive meaningful dimensionless groups is presented (Ruark 1935).

Step 1	Identify the governing differential equations, boundary conditions, and or initial conditions that describe the problem.
Step 2	Transform variables into general dimensionless space.
Step 3	Substitute the general transformations into the differential equations.
Step 4	Multiply the differential equations by the selected scaling factor.
Step 5	Identify the scaling groups from the general differential equations.
Step 6	Reduce the number of groups while maintaining the form of the equations, e.g., set group equal to zero or an arbitrary constant such as one.
Step 7	Identify the remaining dimensionless groups that satisfy the scaling requirements for the problem.
Step 8	Analyze dimensionless groups for dependency and reduce if necessary.

In experimental design, the model and prototype should maintain similarity (same ratios) and be homologous (point-to-point correspondence) so that the mathematical operations performed on them obey mathematical rules. Three levels of similarity are often observed in experimental design. Geometric similarity (required for kinematic similarity) maintains the same ratio of all lengths, giving the model and prototype the same shape. Kinematic similarity maintains the same ratio of all lengths and times, subjecting the model and prototype to velocities of the same ratios. Dynamic similarity maintains the ratio of all forces between the model and the prototype. Unscaled laboratory experiments may negatively impact field operations because some variables in laboratory experiments may be unduly magnified or suppressed (Doscher and Lechtenberg, 1980). Proper scaling is necessary to model phenomena such as fluid flow or phase behavior in confined pore space at micro- to nano-pore scale. A principal challenge to upscaling techniques for multi-phase fluid dynamics in porous media is understanding which properties on the

micro-scale can be used to predict the flow and spatial distribution of phases at the core- and field scales (Pyrak-Nolte et al., 2004). Leverett et al. (1942) recognized the importance of dimensionally scaled models of elements in oil fields to overcome complicated mathematics and erroneous conclusions.

Peters' (2012) emphasizes two critical aspects of dimensionless variables in dimensional analysis:

1. Comprehensiveness: This principle ensures that the chosen dimensionless groups account for all physical influences relevant to the system. The Buckingham Π Theorem is typically used to systematically derive these groups, ensuring that no significant variable is omitted (Peters, 2012). In this study, comprehensiveness is maintained by incorporating key parameters such as treatment pressure, proppant volume, fluid viscosity, and reservoir depth when constructing the dimensionless predictors. These predictors collectively represent the essential physical processes governing hydrocarbon production, avoiding an incomplete representation of the system.

2. Exclusiveness: Exclusiveness ensures that the dimensionless groups are independent of one another and do not introduce redundant information. This prevents collinearity in machine learning models, improving computational efficiency and interpretability (Peters, 2012). In this study, a minimal set of dimensionless groups is selected that provide distinct physical insights without overlapping in meaning. For instance, while both Π_1 (hydraulic stimulation force) and Π_2 (hydraulic stimulation volume) relate to completion characteristics, they capture different aspects of the process, ensuring exclusiveness.

By integrating these principles, the proposed methodology ensures that machine learning models trained on these dimensionless predictors maintain physical interpretability, reducing reliance on purely statistical feature selection. This approach aligns with Peters (2012) work, which advocates for rigorously derived dimensionless groups to enhance model generalizability and predictive reliability.

Acknowledging the fundamental distinctions between reduction methodologies through nondimensionalization and utilizing latent variables via data mining and artificial intelligence is imperative. Firstly, nondimensionalization aims to elucidate the intrinsic relationships among problem attributes through DA without disregarding the essential characteristics of problem representation. This approach generates dimensionless groups that encapsulate relationships independent of the magnitudes of descriptive attributes' numerical values. Conversely, the latter method amalgamates attributes based on their numerical correlations, rendering it a purely statistical exercise. Secondly, the dimensionless groups derived from DA typically possess physical significance, representing variable ratios of identical dimensions. This contrasts the latent variables derived from statistical analysis, which often lack physical representation. Hence, dimensionless groups from DA can provide insights into the 'why' behind observed phenomena, a capacity not inherently possessed by latent variables. Thirdly, dimensionless groups established through DA maintain their independence from the measurement scale, whereas latent variables may strongly depend on the specifics of sample collection and algorithmic selection of sampling. Despite these differences, the methodologies are not mutually exclusive. A synergistic approach can be employed where DA is conducted initially, followed by applying artificial intelligence algorithms to

refine the variable set further. This integrative strategy leverages the strengths of dimensional reduction and data mining techniques, enhancing the analytical capacity for complex problem-solving (Mask et al., 2025)

Dimensionality reduction techniques bolster efficiency and decrease run time. Alghazal and Krinis (2023) underscore the efficacy of Principal Component Analysis (PCA) and recursive feature elimination (RFE) in enhancing prediction accuracy, as evidenced in the RFE ML model's reduction of predictors from over 20 to 7 optimal features. As described by Li et al. (2021), dynamic production rescaling entails a nondimensionalization technique applied to the input predictor cumulative production over cumulative production to forecast cumulative production. However, while DRP addresses the scale of the input predictor, it does not alleviate the size of the hyper-comprehensive input matrix encompassing geographic, geologic, wellbore, well spacing, and completion data. In contrast, the dimensional analysis method proposed in this study offers a distinct approach. By employing dimensional analysis, the input matrix can be simplified by deriving dimensionless predictors from completion and reservoir characterization data. Mask et al. (2019) elucidated how Buckingham Π Theorem could derive dimensionless groups (Reynolds number, Froude number, Eötvös Number) with physical meaning associated with fluid flow patterns. These derived dimensionless groups were predictors in an ML model to predict flow patterns, achieving an accuracy of over 90% (Mask et al., 2019). Similarly, Ojeda (2023) utilized DA to transform input predictors to dimensionless features, revealing the influence of parameters like Weber and gas Reynolds numbers on downhole separator performance and facilitating ML model efficiency. Rahmanifard et al. (2022) emphasize the imperative of comparing computational demands between ML

models and statistical models. This underscores the significance of reducing input predictors to improve efficiency and save time, a significant consideration in addressing complex challenges faced using ML models to forecast hydrocarbon production.

2.1.2. Transfer Learning

Transfer learning involves leveraging knowledge gained from a related task to facilitate learning a new task using past learning experience to expedite the acquisition of novel, yet related, concepts compared to beginning from nothing (Yang et al., 2013). It enables models to generalize more effectively, particularly in scenarios where labeled data is scarce or expensive to obtain. Transfer learning is often applied to deep learning models with a positive, negative, or neutral transfer of knowledge, a concept pioneered in the early 1970s (Bozinovski, 2020). The approach allows models trained on data-rich domains to be fine-tuned for applications in data-limited environments, reducing the need for extensive new training while improving prediction accuracy.

Cornelio et al. (2023) emphasize that not all knowledge is transferable; incorrect transfer may hinder retraining and lead to inaccurate predictions. Cornelio et al. (2023) presented a data-driven approach to rank source datasets for transfer to a target dataset with limited training data available and coupling physics-based models or analytical equations that can be reliably used to transfer knowledge. They emphasized the importance of selecting relevant knowledge to avoid negative transfer effects in deep learning models applied to hydrocarbon production forecasting.

Cornelio et al. (2023) examined the role of proper data selection in transfer learning, particularly for predicting the production performance of undrilled wells. They underscored

the risk of negative transfer, which occurs when knowledge from a source dataset is improperly applied to a target dataset, leading to reduced model accuracy. To mitigate this risk, they introduced a framework for ranking source models based on their relevance to the target dataset. Their findings show that when selecting multiple source models for transfer learning, it is crucial to quantify the domain similarity between source and target datasets to ensure only meaningful knowledge is transferred (Cornelio et al., 2023).

While there is limited research using transfer learning in hydrocarbon production forecasting, studies have shown that data from analogous fields can significantly improve ML production forecasts (Odi et al. ,2021; Cornelio et al., 2023; Mask and Wu, 2023; Misra et al., 2024). This study demonstrates that selecting data from a single adjacent field or multiple fields can positively or negatively impact cumulative production forecasts, highlighting the need for careful dataset selection. Misra et al. (2024) further explained how transfer learning could be used as a tool for production forecasting where data is limited by leveraging knowledge from previously studied reservoirs to refine predictions for new wells. Additionally, transfer learning can improve model generalization by adapting features learned from mature reservoirs to emerging fields, helping overcome the challenge of data sparsity in unconventional plays. By optimizing the selection of source datasets, transfer learning enables more reliable long-term production forecasting while reducing computational costs associated with training models from scratch.

Falola et al. (2023) applied transfer learning to predict CO_2 saturation distribution and pressure plume behavior under uncertain geological and operational conditions. By integrating Fourier Neural Operator (FNO) with transfer learning, they significantly reduced computational time and data storage requirements while maintaining prediction

accuracy. This approach is particularly important for carbon sequestration modeling, where high-resolution simulations are computationally expensive. The findings demonstrate that leveraging pre-trained models on related datasets enables faster and more efficient forecasting, making real-time assessment of CO_2 storage viability more feasible in large-scale subsurface applications.

Misra et al. (2024) explored how transfer learning could be employed as a solution for production forecasting when data availability is constrained. Their study assessed multiple scenarios, demonstrating that when fewer than 100 training realizations are available, transfer learning outperforms conventional ML approaches by leveraging pre-trained models to enhance generalization. However, the study also highlighted that as training dataset sizes increase beyond 500 realizations, transfer learning offers diminishing benefits and may not be necessary. The study demonstrated how ML models trained on source datasets with extensive geological and completion parameters, including porosity, permeability, and fracture characteristics, could be effectively transferred to target datasets with limited production history, improving predictive performance (Misra et al., 2024).

Odi et al. (2021) applied transfer learning for production forecasting in shale reservoirs, demonstrating that ML models trained on data-rich regions, such as La Salle County in the Eagle Ford shale, could be successfully transferred to predict production in less-developed areas with limited data, including Frio, McMullen, Webb, and Dimmit counties. The study highlights the importance of identifying key reservoir performance drivers, such as fracture half-length and gas production rates, to ensure that transferred models retain predictive accuracy in different geological settings. Their results show that transfer

learning significantly reduces uncertainty in production forecasting by leveraging information from established reservoirs.

2.1.3. Fluid Flow Pattern Recognition

Flow pattern recognition is a fundamental aspect of multiphase flow analysis, particularly in oil and gas systems where gas and liquid phases coexist in reservoirs, wellbores, or pipelines. The interaction between these phases gives rise to distinct flow regimes, such as stratified, slug, annular, and churn flow, each with unique characteristics that influence pressure drop, liquid holdup, and flow stability. Accurately identifying these regimes is essential for the design and operation of pipelines, artificial lift systems, and surface processing facilities.

Traditional flow pattern prediction has relied heavily on empirical correlations and mechanistic methods derived from experimental studies. Pioneering works such as Baker (1954), Mandhane et al. (1974), and Taitel and Dukler (1976) introduced flow regime maps that relate flow patterns to superficial gas and liquid velocities across varying pipe orientations. These maps served as practical tools for visualizing regime boundaries and guiding pipeline design.

To improve physical fidelity, later models incorporated mechanistic approaches. For example, Barnea et al. (1982) developed models based on conservation of mass, momentum, and energy, focusing on interfacial forces and phase stability. Arya and Gould (1974) identified discontinuities across flow regime boundaries, highlighting the complexity of modeling transitions between regimes.

While empirical models are fast and easy to apply, they are often limited in scope and applicability. Mechanistic models, on the other hand, offer more physical interpretability and adaptability to varying conditions by solving governing equations that represent the underlying physics. However, they typically require detailed input parameters, such as accurate fluid properties and pipe geometry, and are sensitive to calibration and computational load. Moreover, they can struggle with modeling smooth transitions between regimes and may fail under highly variable conditions.

Despite these limitations, mechanistic models form an important bridge between empirical methods and modern data-driven techniques, especially when experimental data is limited but physical principles are well understood. **Table 5** summarizes key contributions in this field, from early flow regime maps (e.g., Baker, 1954; Mandhane et al., 1974) to modern computational and AI-driven models (e.g., Li et al., 2014; Mask et al., 2019), highlighting the progression of flow pattern prediction methodologies. Machine learning techniques have improved accuracy in complex flow conditions, such as liquid loading in gas wells and highly viscous multiphase systems.

Table 5. The evolution of two-phase flow modeling is summarized, highlighting the transition from empirical flow regime maps to AI-driven predictive models (Baker, 1954; Beggs and Brill, 1973; Gould et al., 1974; Mandhane et al., 1974; Taitel and Dukler, 1976; Barnea et al., 1982a; 1982b; Li et al., 2014; Brito et al., 2015; Al-Safran and Al-Qenae, 2018; Mask et al., 2019).

Author	Year	Study
1954	Baker	Presents a flow regime map for two-phase flow patterns.
1973	Beggs and Brill	Developed multiphase flow correlations in pipelines.
1974	Gould et al.	Established a model to predict pressure distribution in two-phase flow through vertical and inclined pipes.
1974	Mandhane et al.	Generate a flow regime map for horizontal flow using superficial liquid and gas velocities as coordinates.
1976	Taitel and Dukler	Proposed a flow regime map to predict the flow regime transitions in horizontal and nearly horizontal two-phase flow.
1981	Arya and Gould	Compared two-phase liquid holdup and pressure drop correlations across flow regime boundaries. The study indicated that discontinuities exist across flow regime boundaries.
1982	Barnea et al.	Presented mechanistic models to predict flow pattern transitions in inclined pipes.
2014	Li et al.	Utilized Support Vector Machine models to identify flow patterns in gas-liquid flow. The SVM model was deployed to identify liquid loading problems in gas wells.
2015	Brito et al.	Developed a model to forecast flow patterns of gas-highly viscous fluids in horizontal pipelines.
2018	Al-Safran and Al-Qenae	Presented transition models to predict flow pattern transitions in gas-highly viscous oil flow in horizontal pipes.
2019	Mask et al.	Developed a machine learning model utilizing dimensionless predictors to identify fluid flow patterns.

Although traditional models remain useful, they often fail to generalize outside their intended design conditions. Studies such as Kang et al. (2002) and Oddie et al. (2003)

showed that flow regime transitions in large-diameter pipelines differ significantly from predictions based on smaller-scale experiments. Omebere-Iyari et al. (2007) and Abduvayt et al. (2003) demonstrated how high-pressure and inclination effects can alter flow patterns in ways not captured by standard models.

Fluid viscosity also plays a major role. For example, Gokcal et al. (2010) and Khaledi et al. (2014) modified mechanistic models to better predict flow patterns in high-viscosity oil-gas systems, while Vieira et al. (2018) provided further insights on how viscosity impacts flow transitions in upward-inclined pipes. These findings highlight the need for adaptable models that can handle diverse fluid properties and complex pipeline geometries.

Recent developments in machine learning have introduced powerful alternatives for flow pattern prediction. Unlike traditional models, ML algorithms can capture nonlinear, high-dimensional relationships between input features and flow regimes. For example, Li et al. (2014) applied SVM models to improve classification accuracy in gas-liquid flows, particularly for liquid loading in gas wells. Mask et al. (2019) demonstrated how dimensionless predictors, when integrated into ML models, enhance generalization and physical consistency across varying operational conditions.

These modern approaches offer significant improvements in predictive capability, especially for complex, real-world systems where geometry, fluid properties, and operating parameters interact in non-obvious ways. The integration of physics-informed features into ML models holds promise for more robust, generalizable, and interpretable flow regime classification.

2.1.4. Hydrocarbon Production Forecasting

Hydrocarbon production forecasting encompasses a wide array of methodologies, ranging from detailed reservoir simulations with history matching to more streamlined empirical techniques. This diversity allows engineers and decision-makers to balance the need for modeling accuracy with the resources available, which is especially relevant in early-stage developments of tight oil and shale gas reservoirs. For multi-fractured horizontal wells, accurate EUR and production performance forecasts are vital in assessing economic viability, planning well maintenance, and optimizing drilling schedules.

Unconventional reservoirs present additional forecasting challenges due to their inherent complexity. Factors such as heterogeneous rock properties, multifaceted drilling and completion strategies, varied artificial lift systems, and dynamic surface operations can cause large fluctuations in observed well performance over time. Childers and Wu (2020) depict in detail the changes in flow regimes encountered in the reservoir during the life of a shale gas well seen in **Figure 1**. The intricate pore structure, multi-scale fracture network, and coupled flow mechanisms make building a reliable reservoir simulation model for unconventional resources problematic. High-accurate finite difference and numerical integration methods can be adopted to improve fluid flow calculations by minimizing the error in direct permeability calculations (Wang and Sun, 2016). Wang and Sun (2016) emphasized that without careful control of numerical errors, the results may be highly distorted.

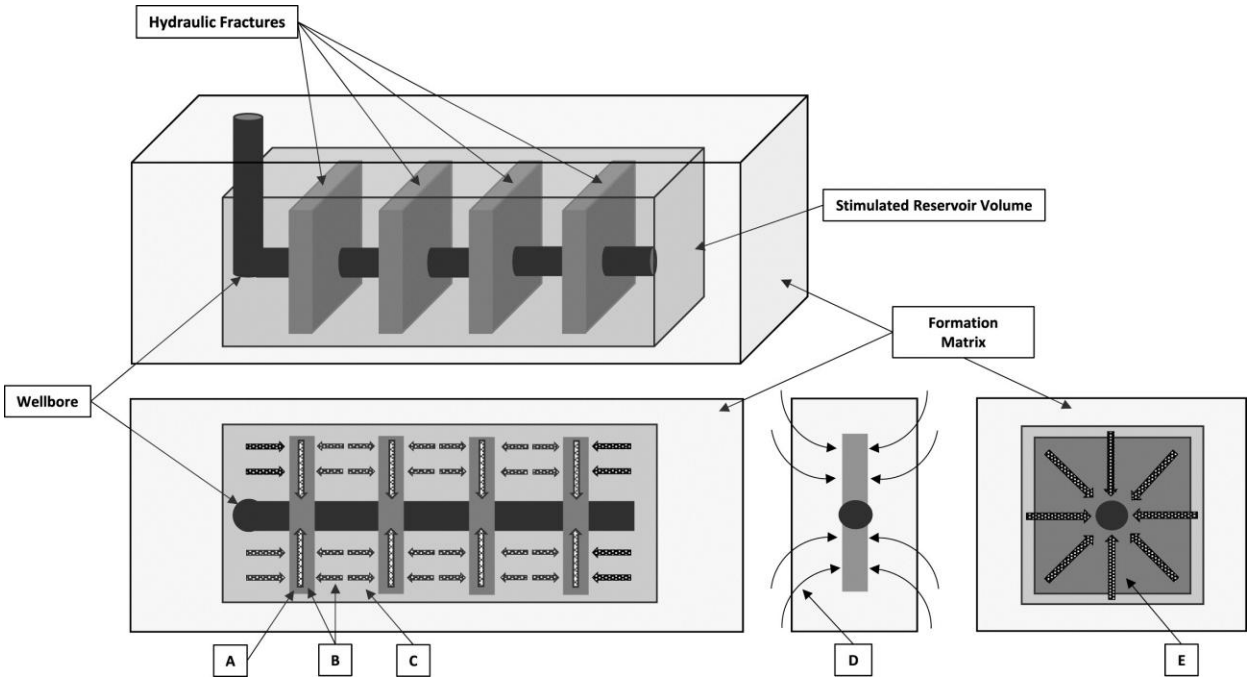

Figure 1. Schematic of flow regime phenomena during shale gas production. Point A depicts fracture linear flow, B depicts bilinear flow, C depicts formation linear flow, D depicts the transition from linear flow to radial flow, and E depicts pseudo-radial flow (Childers and Wu, 2020).

Considering these complexities, machine learning has gained traction as a robust tool for predicting production trends and EUR. ML-based forecasting, often referred to as ML-DCA (Stephenson, 1999; Cao, et al. 2016; Kamari, et al. 2017; Li and Han, 2017; Vyas et al., 2017; Amr et al., 2018; Sarduy and Edelmann, 2018; Bakay et al., 2019; Dupont et al., 2020; Li et al., 2021; Liu et al., 2021; Odi et al., 2021; Gu et al., 2022; Jha et al., 2022; Li et al., 2022; Cornelio et al., 2023; Mask and Wu, 2023; Vega-Ortiz et al., 2023) can efficiently capture nonlinearities in production data. However, ML-DCA also demands careful management of potentially large input matrices and hyperparameter tuning (Mask and Wu, 2023).

When determining the best approach for production forecasting, engineers must carefully consider each method's strengths and drawbacks. Physics-based reservoir models can yield highly detailed insights but are computationally intensive and require extensive subsurface data. Empirical models, while generally simpler to implement, may overlook key reservoir intricacies. Machine learning methods offer a strong balance between these two extremes, provided the complexity of the input space is well managed. As unconventional resource development continues to expand, refining and integrating these forecasting techniques will be paramount to maintaining efficient and cost-effective operations.

Probabilistic forecasting in hydrocarbon production incorporates uncertainty quantification by using statistical models, Monte Carlo simulations, and Bayesian approaches to estimate a range of possible production outcomes rather than a single deterministic forecast. This methodology accounts for geological variability, well performance fluctuations, and operational uncertainties by producing probability distributions of expected production. Unlike deterministic models, probabilistic forecasting enables decision-makers to assess risks, optimize resource allocation, and improve economic evaluations by considering best-case, worst-case, and most-likely scenarios. When integrated with decline curve analysis and physics-based models, probabilistic methods allow engineers to better quantify uncertainties inherent in unconventional reservoirs and multi-fractured horizontal wells, resulting in more informed and resilient forecasting strategies (Williams-Kovacs and Clarkson, 2011; Morales-German et al., 2012).

Reservoir models serve as a robust tool for production forecasting. Rate transient analysis (RTA) can be used to analyze production data to determine reservoir performance and characteristics. A reservoir simulation model can be constructed to forecast future production based on the reservoir characteristics acquired from RTA (Mishra, 2014; Mireault and Dean, 2008). However, their efficacy relies heavily on comprehensive reservoir characterization encompassing core data, fluid properties, well logs, and production records. While numerical simulations hold the potential for highly accurate predictions, capturing the physics of unconventional reservoirs remains challenging, especially regarding gas storage, fluid flow, and transportation mechanisms (Mohaghegh and Abdulla, 2014).

The production-type curve is obtained by trending the average of the existing wells' production history (Mask and Wu, 2023) often using Arps equation. Production-type curves are most often created from well data by normalizing the production time of the like kind wells, averaging the production rate, and generating a production forecast using DCA. A production-type or cumulative production curve presents a well production profile for a specific geographical location, drilling/completion design, and formation with similar reservoir characteristics. Engineers must consider many factors when forecasting production-type curves for MFHW in unconventional reservoirs. MFHW production-type curves require attention to the initial production period as the hydraulic stimulation creates an artificially inflated transient period where the production can increase initially and then flatten before the actual decline period begins. Choke management also plays into how the initial production profile behaves. The drawdown of the well also affects how the proppant seats. Drawing the well down too quickly can lead to crushed proppant,

proppant embedment leading to reduced conductivity and a steeper decline profile. The order in which the wells are drilled affects the production profile. The first well will likely have the highest production given that the reservoir pressure has not been disturbed, hydrocarbons have not been depleted, and neighboring wells are not causing interference.

Arps model (1945) and its derivatives are widely used in the oil and gas industry for production forecasting, utilizing historical production data to predict future performance. Arps uses past performance to mathematically predict future performance under the assumption that operating conditions are constant, drainage area remains constant, bottom hole pressure, and no change in skin.

To maintain consistency with the dimensionless framework used in this study, Arps' equations can be rewritten in a dimensionless form, offering a more compact representation and enabling better comparison across diverse reservoir conditions. Arps equation for exponential rate in dimensionless form:

$$q_D = e^{-t_D} \quad (1)$$

The hyperbolic decline model in its dimensionless form is expressed as:

$$q_D(t_D) = (1 + bt_D)^{-1/b} \quad (2)$$

Arps loss-ratio (dimensionless form):

$$\frac{1}{D_D} = \frac{q_D}{\frac{dq_D}{dT_D}} \quad (3)$$

where:

$q_D = \frac{q}{q_i}$ is the dimensionless production rate,

$t_D = D_i t$ is the dimensionless time,

$D_D = \frac{D}{D_i}$ is the dimensionless decline rate,

$b = -\frac{d}{dt}\left[\frac{q}{\frac{dq}{dt}}\right]$ is the hyperbolic exponent.

The hydrocarbon production system is a complex system impacted by multi-phase fluid flow that requires frequent adjustments to the production network to maintain operations (Xue 2024). However, the Arps model's premise of constant operating conditions often diverges from reality, particularly in unconventional multi-fracture horizontal wells where critical factors like drainage area, bottom hole pressure, and well productivity index are inherently dynamic. Also, empirical models struggle to capture the complex flow regimes, transient behaviors, and fracture interference patterns observed during the production of MFHW in unconventional resources. Engineers often resort to curve-fitting techniques, potentially overlooking critical model assumptions to achieve short-term forecast accuracy (Mask and Wu, 2023).

Despite its widespread use, Arps' model falls short in capturing the complexities of MFHW production due to several key limitations. First, unconventional reservoirs exhibit prolonged transient flow behavior resulting from complex fracture networks and extremely low permeability, conditions that deviate significantly from the steady-state assumptions underpinning Arps' equations. Additionally, fracture interference caused by well-to-well interactions in tightly spaced completions can lead to deviations in production performance that the Arps model is not equipped to handle. Another major challenge is

the dynamic nature of the drainage area in unconventional plays, which expands over time as pressure depletes and more of the reservoir is accessed. This behavior violates the model's assumption of a constant drainage area. Furthermore, bottom-hole pressure (BHP) in unconventional wells often fluctuates due to operational adjustments or changing reservoir conditions, undermining the Arps model's assumption of constant flowing pressure. These limitations collectively reduce the model's reliability and accuracy for forecasting production in unconventional reservoirs, particularly during early-time and transient-dominated periods.

Many decline curve analysis models have been developed to capture the flow regime transitions observed during multi-fractured horizontal well production in unconventional reservoirs. The dimensionless forms of these equations are presented in **Appendix C.** Ilk et al. (2008) introduced the Power Law Exponential (PLE) decline model, which accounts for the transition from transient to boundary-dominated flow by incorporating a power-law function into the exponential decline equation. The advantage of the PLE model lies in its ability to capture both early- and late-time flow behaviors. In the early time, the t^n term effectively models the transient flow regime, while at late times, the decline is governed by D_∞, ensuring that production rates asymptotically approach zero (Ali and Sheng, 2015). However, despite its flexibility, the PLE model introduces additional empirical parameters that lack clear physical interpretation, making it difficult to directly relate model outputs to reservoir characteristics. Its accuracy is highly dependent on high-quality production data, and fitting the model can be sensitive to noise, particularly when late-time data are sparse. As such, while the PLE model enhances forecasting range, it

provides limited physical insight and may not generalize well across wells with highly variable flow behavior.

The Stretched Exponential Decline (SEPD) model, introduced by Valkó (2009), is a physics-based approach that uses governing differential equations to characterize decline behavior in unconventional reservoirs. Unlike traditional decline models, SEPD accounts for reservoir complexity by modeling production as the aggregate response of multiple independently decaying subsystems. In this formulation, t represents actual production time, while τ denotes a characteristic time scale that standardizes time across varying reservoir conditions. This transformation allows the SEPD model to generalize decline trends more effectively and improve predictive accuracy across different reservoir settings. The model is grounded in the premise that a single production system can be represented as a summation of multiple decaying processes, each governed by two-parameter gamma functions (Franchi, 2020; Valkó and Lee, 2010). This structure provides a flexible framework capable of capturing the complex, multi-rate behavior typical of multi-fractured horizontal wells in unconventional plays. However, despite its improved accuracy, the SEPD model remains largely empirical in nature and lacks direct physical parameters that can be easily interpreted or calibrated using standard reservoir data. Its reliance on curve-fitting also introduces potential risks of overfitting, particularly in datasets with high noise or limited production history.

Duong's (2010) model assumes that production is dominated by fracture flow, with negligible contribution from the matrix. It postulates that connected fracture density increases over time as local stress changes, induced by pressure depletion, reactivate existing fractures and potentially compromise the hydraulic integrity of shale seals. Duong

demonstrated that a log-log plot of cumulative production versus time exhibits a unit slope, regardless of fracture type, enabling simplified interpretation of production behavior. The model characterizes MFHW performance in unconventional reservoirs by analyzing the slope and intercept of this log-log relationship, offering insights into reservoir properties, completion design, and flow mechanisms. However, a key limitation of the Duong model is its assumption that matrix contribution is negligible, which may not hold true in reservoirs where dual-porosity systems significantly influence long-term production. Furthermore, because the model is derived empirically and lacks a direct physical foundation, its predictive accuracy diminishes over extended production periods, particularly as flow transitions from fracture-dominated to matrix-supported regimes.

A comparative study by Yehia et al. (2022), summarized in **Appendix D**, evaluated the performance of 20 widely used decline curve models, highlighting the strengths and weaknesses of each under varying data conditions. The study revealed that model accuracy in both goodness-of-fit and forecast reliability was highly sensitive to data availability and outlier treatment strategies. **Figure 2** illustrates the impact of data sparsity, showing how forecasting with only 20% of production history can lead to significant deviations across different models. In contrast, **Figure 3** demonstrates that fitting 80% of the available data reduces forecast variability and improves alignment with actual production. This finding underscores the critical role of data volume in ensuring reliable long-term forecasts, an essential consideration for early-stage asset evaluation, economic planning, and operational decision-making. In unconventional reservoirs, where early-time production is often used to project long-term performance, insufficient data can lead to overoptimistic forecasts and misinformed development strategies.

Therefore, understanding how data availability affects model accuracy is vital for reducing risk and improving the confidence of forecasting outcomes.

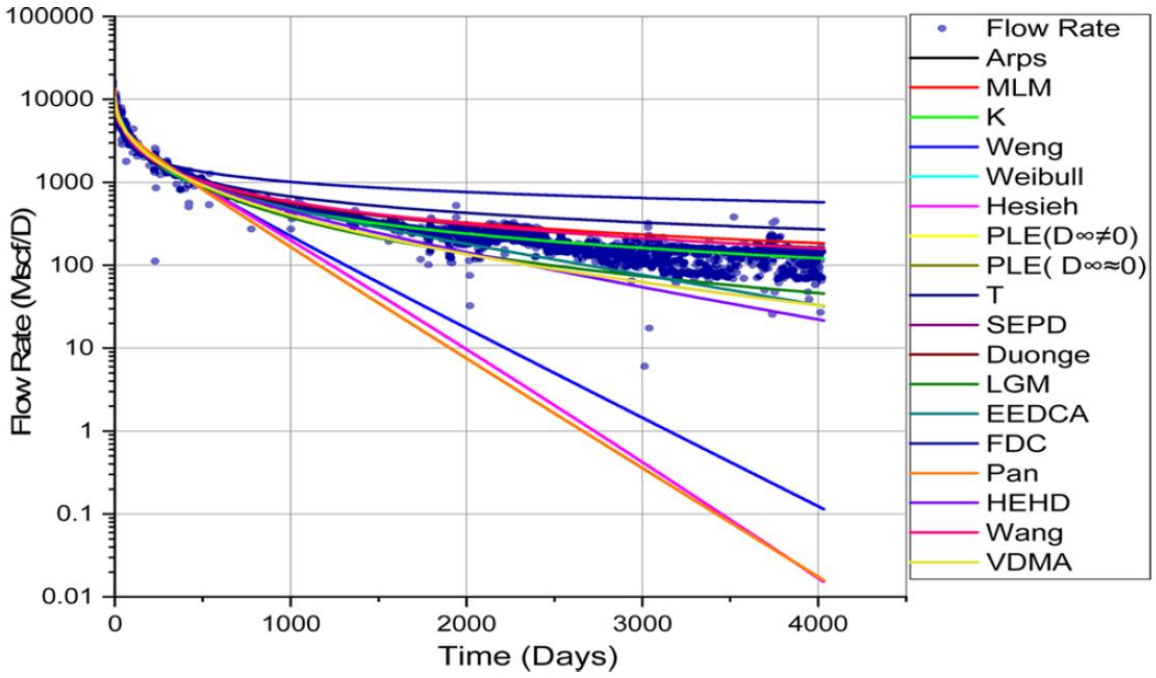

Figure 2. The variability in production forecasts generated by different decline curve models when trained on only 20% of the available production data. Significant deviations between models highlight the sensitivity of forecasting methods to data availability and model assumptions. The study emphasizes the importance of selecting an appropriate DCA model based on data volume, outlier treatment, and the model's inherent assumptions Yehia et al. (2022).

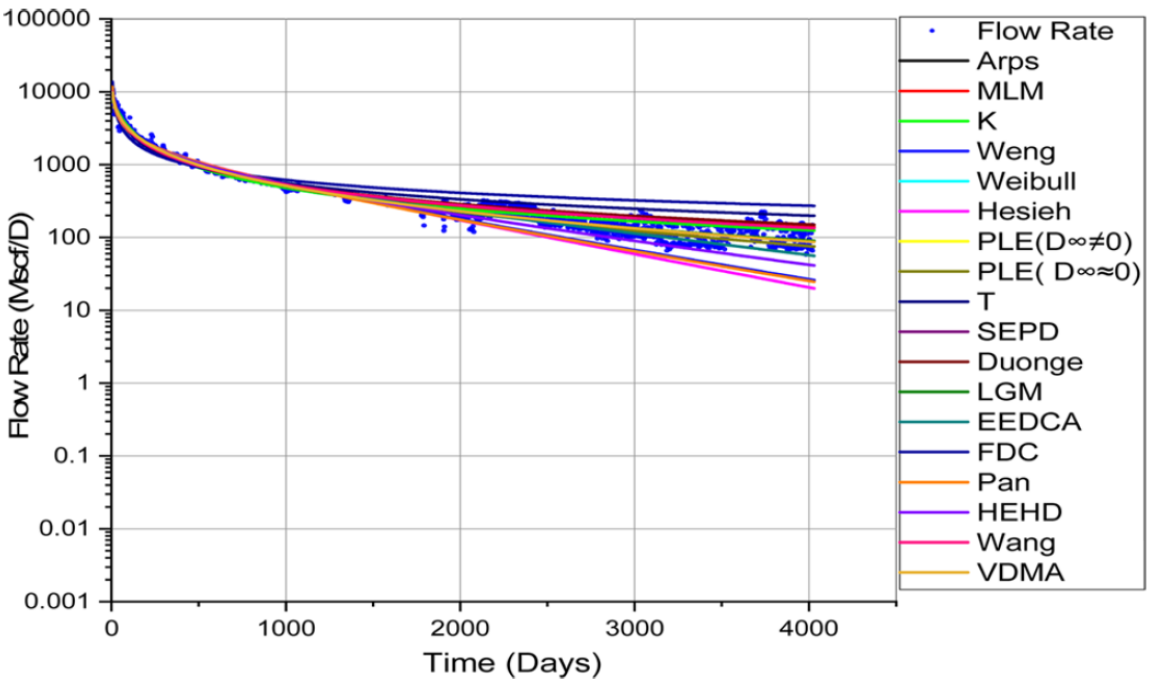

Figure 3. The DCA forecasts improve when fitted with 80% of the production data. As a result, many of the DCA models begin to converge more closely to the actual production trend (blue dots), reducing variability in long-term predictions. This highlights the importance of data availability in enhancing the reliability of production forecasts, as shown in Yehia et al. (2022).

Ali and Sheng (2015) evaluated the accuracy of rate-time decline models compared to cumulative production-based approaches for forecasting EUR. Their study revealed that rate-time models can overestimate EUR by more than 150%, particularly in unconventional reservoirs where prolonged transient flow dominates. This highlights a key limitation of traditional DCA methods, which often fail to capture long-term production behavior in multi-fractured horizontal wells. In contrast, Ali et al. (2014) proposed a cumulative production decline approach, demonstrating that wells producing at constant flowing pressure exhibit a straight-line slope of unity on a log-log plot of cumulative production versus time. Their findings suggest that cumulative-based models provide a

more reliable and physically consistent forecasting framework, especially for tight and shale gas reservoirs where fracture flow governs production.

Building on this, Ali and Sheng (2015) compared multiple DCA methods and concluded that cumulative production models yield more stable and accurate predictions than rate-time approaches, particularly in fracture-dominated systems. By integrating cumulative production trends rather than relying solely on instantaneous rate decline, the forecasting framework developed in this study addresses the limitations of traditional models, reducing overestimation risk, improving robustness in transient flow regimes, and enhancing generalizability across diverse unconventional reservoirs.

While probabilistic analysis is a fundamental tool in reservoir engineering, its effectiveness is often constrained by data availability. In scenarios with sparse information, Morales-German et al. (2012) proposed generating synthetic reservoir models for shale gas wells using statistical representations of major shale basins, combining decline curve analysis with probabilistic resource estimates. This approach helps address uncertainty during early exploration but depends on access to analogous field data and may fail to capture site-specific reservoir characteristics. Similarly, Williams-Kovacs and Clarkson (2011) demonstrated that Monte Carlo simulation can effectively quantify uncertainty in production forecasting by generating a wide range of potential outcomes. However, in practice, this method presents significant challenges, including the need for accurate input distributions, the difficulty of defining realistic probability functions, and the high computational cost of solving high-dimensional problems with multiple uncertain parameters. These limitations highlight the need for forecasting

frameworks that balance probabilistic rigor with physical realism and computational efficiency, particularly in the context of unconventional resource development.

2.1.5 Machine Learning

Machine learning leverages statistical models and algorithms to predict outcomes from data without requiring explicitly programmed instructions. Implementing a machine learning (ML) model is a process-intensive task that demands a strong foundation in data preprocessing, algorithm selection, and programming. ML algorithms are generally categorized into supervised, unsupervised, semi-supervised, and reinforcement learning. In supervised learning, the model learns a function that maps inputs to outputs using labeled input-output pairs (Sarker, 2021). These supervised algorithms are typically grouped into two problem types: classification, which yields categorical outputs, and regression, which yields continuous outputs (Hastie et al., 2009).

In this study, supervised classification algorithms are employed for fluid flow pattern recognition, as the goal is to identify discrete flow regimes. These algorithms enable accurate categorization of complex multiphase flow behaviors based on input features, enhancing pattern recognition across varying operating conditions. For hydrocarbon production forecasting, a supervised regression algorithm is applied to predict continuous outputs, such as cumulative production over time. Regression models estimate the relationship between input features and output targets, allowing the model to infer production behavior from observed data (Kononenko and Kukar, 2007).

Supervised models are trained by iteratively refining predictions until satisfactory performance is achieved. The goal is to approximate the underlying mapping function

with enough accuracy to generalize effectively to previously unseen data (Hastie et al., 2009; Saleh et al., 2022). During training, the model identifies patterns and correlations in the data, which it then applies to make accurate predictions in real-world scenarios. This generalization capability is critical for deploying models in dynamic operational environments, where variability in inputs and conditions is common. The mathematical formulations of the supervised learning expressions used in this study are provided in **Appendix E**.

1. Linear Regression.

Linear regression is a foundational algorithm used to model the linear relationship between input (independent) variables and a target (dependent) variable. The key assumption is that the conditional expectation of the target variable, $E(Y|X)$, is a linear function of the input features $x_1,\cdots,x_n$ (Hastie et al., 2009). The model estimates the coefficients $\beta_0, \ldots, \beta_n$ by minimizing the sum of squared differences between the observed and predicted values through a least-squares optimization process. When there is only one predictor variable ($n = 1$), the model is referred to as simple linear regression; when there are multiple predictors ($n > 1$), it is known as multiple linear regression.

Linear regression is exclusively suited for regression problems, as it predicts continuous outputs rather than categorical class labels. However, a related model, logistic regression, adapts the linear regression framework for classification tasks by applying a sigmoid function to constrain the output between 0 and 1, making it appropriate for binary and multiclass classification problems (Hastie et al., 2009).

Linear regression performs best under specific conditions: when the relationship between inputs and the target is approximately linear, the dataset is free from strong multicollinearity and influential outliers, and the primary goal is to interpret the effect of each feature on the outcome. Its high interpretability and computational efficiency make it a useful baseline model. However, its limitations become apparent when modeling complex systems, such as non-linear reservoir behavior, where linear assumptions fail unless extended through feature transformations or polynomial terms. In this study, linear regression is used as a baseline model for comparison, providing a reference point to evaluate the performance improvements achieved by more advanced machine learning algorithms.

2. Decision Tree.

A Decision Tree (DT) is a non-parametric, supervised learning algorithm capable of performing both classification and regression tasks. It works by recursively partitioning the input space into subsets based on feature values, forming a hierarchical, tree-like structure (Mitchell, 1997). The goal is to create increasingly homogeneous groups within each partition, thereby improving the model's predictive accuracy.

For classification, decision trees evaluate possible feature splits using impurity measures such as Gini index or entropy, selecting the split that best separates the classes. The process continues until a stopping criterion is reached, such as a minimum number of samples per leaf or a maximum tree depth (Geurts et al., 2006). In regression tasks, decision trees use criteria like mean squared error (MSE) or mean absolute error (MAE) to split data, aiming to minimize variance within each partition and provide more stable predictions in the leaf nodes.

Decision trees are easy to interpret and computationally efficient, which makes them useful for understanding feature importance and decision rules. However, they are prone to overfitting, particularly when deep trees are grown without pruning or regularization. Their predictive performance can also be sensitive to small variations in the training data. To overcome these limitations, ensemble methods such as Random Forests and Gradient Boosting are often used to combine multiple trees, reducing variance and improving generalization (Geurts et al., 2006).

In this study, decision trees serve as an interpretable baseline model for both classification and regression tasks, providing insight into feature influence and serving as a foundation for more advanced ensemble learning techniques.

3. Random Forest Regression.

Random Forest is an ensemble learning algorithm that builds multiple decision trees and aggregates their predictions to improve accuracy and reduce overfitting. It can be applied to both classification and regression tasks (Breiman, 2001). In classification, each tree casts a vote for the predicted class, and the final decision is made through majority voting. In regression, the output is the average of predictions from all individual trees (Breiman, 2001; Pedregosa et al., 2011).

Random Forests introduce two key sources of randomness that enhance generalization: bootstrap aggregation (bagging) and random feature selection. In bagging, each decision tree is trained on a randomly sampled subset of the data with replacement. At each split in the tree, a random subset of features is considered, which helps to decorrelate trees and prevent the ensemble from overfitting to noise in the training data (Hastie et al., 2009).

This dual-randomization strategy increases model robustness and typically results in higher predictive performance.

RF models are particularly effective for handling high-dimensional datasets and are known for their resilience against overfitting, outperforming single decision trees in terms of generalization. They also provide useful insights into feature importance, offering a degree of interpretability. However, they can be computationally intensive, especially for large datasets, and their ensemble nature makes them less transparent than simpler models.

In this study, Random Forests are used as a benchmark ensemble method for both classification and regression tasks, offering a balance between predictive accuracy, robustness, and interpretability across complex petroleum engineering datasets.

4. Extra-Trees Regression.

Extremely Randomized Trees (Extra-Trees or ETR) is an ensemble learning algorithm that builds multiple decision trees with an increased level of randomness to improve model generalization. Like Random Forests, it can be applied to both classification and regression tasks, but it differs in how it selects split points and constructs trees (Geurts et al., 2006).

Unlike conventional decision trees, which determine optimal split points by evaluating a specific criterion, such as variance reduction for regression or Gini impurity for classification, Extra-Trees selects split thresholds at random from a uniform distribution over the feature range. This added randomness enhances tree diversity, reducing variance and improving generalization, particularly in high-dimensional or noisy datasets.

For classification problems, the final prediction is based on majority voting among trees; for regression, predictions are averaged across the ensemble.

ETR offers several benefits over traditional ensemble methods. It reduces variance more aggressively than Random Forests by introducing greater decorrelation among trees. Additionally, it is computationally efficient, as it does not rely on bootstrap sampling (bagging); each tree is trained on the full dataset. However, this benefit comes with a tradeoff—randomized splits may increase model bias, potentially leading to underfitting when more precise split selection is necessary (Geurts et al., 2006).

In this study, Extremely Randomized Trees are employed as a robust ensemble model for both classification and regression tasks. Their strong generalization capability makes them particularly valuable for modeling complex patterns in hydrocarbon production forecasting, where reducing overfitting is often more critical than achieving minimal bias.

5. Gradient Boosted Decision Tree.

Gradient Boosting is a powerful ensemble learning technique that constructs decision trees sequentially, with each new tree learning to correct the residual errors of the previous ones. This iterative process allows the model to gradually improve predictive performance by focusing on difficult-to-predict instances (Friedman, 2002; Wang et al., 2020). GB is suitable for both classification and regression tasks, offering high accuracy in complex data environments.

In classification problems, GB minimizes a classification loss function such as log-loss, with the final output determined by aggregating predictions through weighted voting. For regression, it aims to reduce error by minimizing a loss function like mean squared error,

updating the model iteratively using gradient descent. At each stage m, a new decision tree is fit to the negative gradient of the loss function with respect to the current model's predictions. The ensemble prediction is updated by adding this new learner to the model, allowing the system to refine its approximation over time (Friedman, 2002). The generalized Gradiant Boosting algorithm can be seen in **Appendix F**.

Unlike Random Forests, which build trees independently and in parallel, Gradient Boosting builds trees sequentially, where each iteration attempts to optimize performance by correcting the shortcomings of its predecessor. This results in a more refined model capable of capturing subtle patterns and nonlinear relationships in the data.

Gradient Boosting offers several advantages, including superior predictive accuracy and the ability to handle complex, nonlinear data distributions. However, it requires careful tuning of hyperparameters such as learning rate, tree depth, and the number of trees to prevent overfitting. The sequential nature of the algorithm also makes it more computationally expensive compared to other ensemble methods, particularly on large datasets.

In this study, Gradient Boosting is utilized to enhance predictive modeling for both flow pattern classification and hydrocarbon production forecasting. Its ability to model intricate relationships and adapt to structured, high-dimensional data makes it well-suited for addressing the complexities of unconventional reservoirs and multiphase flow behavior.

6. Support Vector Machine.

Support Vector Machines (SVM) are a class of supervised learning algorithms primarily designed for classification tasks but can be extended to regression through Support

Vector Regression (SVR) (Vapnik, 2000; Platt, 2000). In classification, SVMs aim to identify the optimal hyperplane that maximizes the margin between different classes, offering robustness against overfitting, especially in high-dimensional feature spaces. For regression tasks, SVR introduces the concept of an epsilon-insensitive loss function, which tolerates small deviations from the predicted output without incurring a penalty. Instead of minimizing the total prediction error, SVR focuses on finding a function that fits the data within a predefined margin, prioritizing generalization over exact matching (Platt, 2000; Vapnik, 2000; Chang and Lin, 2022).

The regression function in SVR is determined by a subset of training data points known as support vectors, which lie closest to the decision boundary. These support vectors are critical in defining the shape and orientation of the regression model. Unlike traditional regression models, SVR offers a more flexible and theoretically grounded approach to capturing complex, nonlinear relationships by utilizing kernel functions such as linear, polynomial, or radial basis function (RBF) kernels. However, its performance is highly sensitive to hyperparameter tuning, particularly the selection of the kernel type, the regularization parameter $C,$ and the epsilon ε margin (Vapnik, 2000; Platt, 2000 Hastie et al., 2009; Chang and Lin, 2022). The details of the SVR algorithm can be seen in **Appendix G**.

Despite being computationally intensive, especially for large datasets due to the underlying quadratic programming optimization, SVR excels in scenarios where the number of features exceeds the number of samples or where nonlinear patterns dominate. In the context of this study, SVR is applied as a regression technique for hydrocarbon production forecasting. Its ability to model complex input-output

relationships with strong generalization makes it particularly suitable for unconventional reservoir systems, where data irregularities and high-dimensionality often limit the effectiveness of traditional models.

7. Neural Networks.

Artificial Neural Networks (ANNs) are a class of machine learning models inspired by the structure and function of biological neurons (McCulloch and Pitts, 1943; Peter et al., 2019). These models are capable of learning complex, nonlinear relationships between input and output variables through a network of interconnected nodes (neurons), making them suitable for both classification and regression tasks. Among the most widely adopted ANN architectures is the Multi-Layer Perceptron (MLP), a feedforward neural network that processes information from the input layer through one or more hidden layers to the output layer (Rosenblatt, 1958; Sarker, 2021). MLPs are trained by minimizing a loss function, such as mean squared error for regression or cross-entropy for classification, using backpropagation and gradient descent algorithms to iteratively adjust network weights and reduce prediction error (Vapnik, 2000).

In contrast to ANNs, Recurrent Neural Networks (RNNs) are specifically designed for sequential and time-dependent data. RNNs retain a hidden state that captures information from previous time steps, enabling the modeling of temporal dependencies (Elman, 1990). This makes RNNs particularly effective for time-series forecasting, where the order and continuity of data are critical. However, training deep RNNs can be challenging due to issues like the vanishing gradient problem, often addressed using advanced architectures such as Long Short-Term Memory (LSTM) networks.

Neural networks offer several advantages for production forecasting. ANNs are highly flexible and excel at capturing complex, nonlinear interactions between variables, making them powerful tools for modeling unconventional reservoir behavior. RNNs, in particular, are effective in capturing sequential dependencies inherent in production data over time. Nevertheless, these models require substantial computational resources and large training datasets to achieve reliable performance, and they can be sensitive to overfitting and hyperparameter selection.

In hydrocarbon production forecasting, machine learning models such as ANNs and RNNs have shown promising results. Amr et al. (2018) demonstrated that ML-based forecasts could outperform traditional Arps decline curve estimates by 20%–30%, utilizing more than 20 predictors encompassing geospatial, geological, petrophysical, drilling, and completion parameters. Similarly, Bakay et al. (2019) developed an ML framework that integrated spatially varying decline curves, k-means clustering for regional categorization, support vector machines (SVM) for uncertainty quantification, and classification and regression trees (CART) to identify key factors influencing production. These applications highlight the growing role of advanced neural network models in improving prediction accuracy, reducing uncertainty, and enhancing decision-making in the oil and gas industry.

Jha et al. (2022) introduced a workflow that integrates statistical and ML methods to automate the application of the multi-segment Arps decline model for unconventional reservoirs. Given the challenge of managing production forecasts for hundreds of wells in shale plays, their approach streamlines forecasting by automating key steps such as outlier detection, flow regime identification, and decline parameter estimation. By

incorporating physics-based constraints, their method enhances both predictive accuracy and efficiency, particularly for large datasets. This study highlights the advantages of integrating data-driven techniques with traditional decline curve analysis, improving reliability and scalability.

Gu et al. (2022) explored deep learning models for gas reservoirs, comparing Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) models. LSTM proved the most accurate, outperforming Duong's empirical forecast. Vega-Ortiz et al. (2023) further demonstrated LSTM's ability to address limitations in empirical models, particularly their reliance on high-quality data and inability to capture flow regime changes. They advocate for coupling deep learning models with economic factors, such as future oil prices and operational costs, to enhance their applicability.

While ML models significantly improve production forecasting, their effectiveness depends on comprehensive training datasets, which are often limited in early-stage unconventional plays. Incorporating completions and reservoir characterization data enhances forecasting accuracy by accounting for well performance and flow regime variations. An early attempt by Stephenson (1999) used neural networks and genetic algorithms to predict hydrocarbon production, integrating drilling, completions, and stimulation data. However, this approach was confined to neural networks and lacked methodological specificity.

Li et al. (2021) introduced an advanced ML-DCA method leveraging dynamic production rescaling (DPR) to forecast production based on historical cumulative production. While DPR achieves high accuracy, it requires production data from the well, limiting its

applicability to undrilled locations. Other ML models focus on predicting decline curve parameters for empirical DCA equations (Karami et al., 2017; Li and Han, 2017; Vyas et al., 2017; Amr et al., 2018; Li et al., 2022). Some approaches integrate physics-based constraints, such as physics-constrained ML (PCML) and physics-informed ML (PIML) (Liu et al. 2021; Jha et al. 2022). While these models offer insight into underlying physics, they assume the governing equations embedded within them are correct.

Many ML-DCA models prioritize selecting the most effective algorithm for production forecasting, frequently favoring neural networks such as RNNs, LSTMs, and GRUs (Gu et al. 2022; Vega-Ortiz et al. 2023). However, training, testing, and validating these models are time intensive. Dimensional analysis (DA) offers a way to reduce the number of input features, streamlining computational requirements while preserving predictive accuracy. By integrating DA with ML models, production forecasting can become more efficient, making ML-driven approaches more practical for real-world applications.

2.1.6 Dataset Preprocessing and Evaluation

Data preprocessing is a crucial step in transforming raw datasets into structured formats suitable for machine learning and statistical analysis. In petroleum engineering applications such as production forecasting and flow pattern recognition, the quality of the input data directly influences model performance and interpretability. Inadequate preprocessing can introduce noise, bias, or inconsistencies that lead to inaccurate or misleading results. This section outlines the preprocessing techniques applied in this study, including the treatment of missing values, outlier detection and removal, normalization, feature transformation, and dataset partitioning. These steps ensure that

the final dataset accurately reflects production behavior in unconventional reservoirs and is optimized for robust and generalizable model training.

Data cleaning is a critical step in ensuring model reliability, especially when dealing with datasets sourced from multiple wells, vendors, or measurement systems. In petroleum engineering, common issues such as missing values, outliers, and inconsistent units can introduce bias and compromise the accuracy of machine learning predictions (Fan et al., 2021; Nicholson, 2018). This study reviews two primary imputation techniques to handle missing data: mean imputation and interpolation.

Mean imputation involves replacing missing values with the average of known values from similar data points. This method is particularly suitable for variables with low variance, as it maintains consistency without significantly distorting the distribution. For example, missing mud weight values were filled using formation-specific averages to preserve geological relevance. Interpolation, on the other hand, estimates missing values by using adjacent data in a continuous sequence. This method is especially effective for time-series production data, as it preserves natural trends and avoids abrupt changes. Linear interpolation can be used to fill short gaps in monthly production records, ensuring a smooth transition between known data points.

For critical reservoir properties like porosity and permeability, where missing values can have a significant impact on model accuracy, analogous formations were used to infer values. Monte Carlo simulations were then applied to introduce variability reflective of real-world reservoir heterogeneity, reducing the risk of overfitting to overly simplistic assumptions.

Outlier detection was essential to ensure that extreme values did not skew model training or performance. A combination of statistical techniques and domain expertise was used. Statistically, outliers were flagged using the Interquartile Range (IQR) and Z-score methods, identifying values that fell beyond 1.5 times the IQR or that deviated significantly from the mean. However, statistical detection alone was insufficient. Each flagged value was validated using engineering domain knowledge to determine whether it represented a true anomaly or an operational outlier. For example, wells with unusually high proppant volumes were cross-checked against completion records. In one case, an initially flagged well was found to reflect a deliberate high-intensity fracture design rather than a data entry error. Rather than removing such data, it was retained and flagged for validation during model training, ensuring the model could learn from valid but rare cases.

These data cleaning strategies helped improve the consistency and representativeness of the dataset. By addressing missing data, correcting outliers, and maintaining geological and operational context, the final dataset better reflects true field behavior. This foundation is essential for building machine learning models that are not only accurate but also generalizable across different reservoir conditions and well designs.

By evaluating both dimensional analysis and PCA independently, this study assesses whether physics-informed dimensional reduction via DA offers superior performance over statistical methods. This comparative approach allows for determining which technique better preserves predictive accuracy, enhances model interpretability, and improves computational efficiency in flow pattern recognition and hydrocarbon production forecasting.

Data scaling is a vital preprocessing step that ensures consistency in feature magnitudes, particularly important for machine learning algorithms that rely on distance metrics, such as SVM and K-Nearest Neighbors (KNN) (Fan et al., 2021). In petroleum engineering applications, input variables, such as pressure, temperature, porosity, and permeability, can span multiple orders of magnitude. Without proper scaling, these differences can distort the learning process, leading to biased models and suboptimal performance.

To address this, multiple scaling techniques were tested based on the characteristics of each variable. Min-Max Scaling was applied to variables like porosity and permeability, constraining them to a [0, 1] range to ensure uniform contribution across features. This approach is particularly effective for features that do not follow a Gaussian distribution and have well-defined bounds. Z-Score Standardization was applied to variables such as pressure and temperature to center the data around a mean of zero with unit variance, which enhances numerical stability for gradient-based algorithms. The mathematical formulations for these scaling methods are provided in **Appendix H.**

In cases where data contained outliers or skewed distributions, robust Z-score standardization was examined, which substitutes the mean and standard deviation with the median and interquartile range, respectively. This method mitigates the influence of extreme values, preserving the integrity of the scaled data.

Practical implementations included rescaling production time to start from zero across all wells to establish a consistent temporal reference. Production volumes were normalized by dividing fluid output by lateral length (e.g., per 1,000 ft) to account for variations in well geometry. Slurry volumes and densities were converted from mass-based to volume-

based units using proppant density, allowing more accurate representation of fluid flow properties.

These scaling procedures were essential to create a numerically balanced input space, enabling the machine learning models to learn more efficiently and generalize more effectively across diverse reservoir and operational conditions.

Data transformation is a crucial step in preparing raw inputs for machine learning models, involving the conversion of data into formats that enhance model interpretability, stability, and performance (Nicholson, 2018). In petroleum engineering, datasets often contain a mix of categorical and continuous variables, as well as non-linear trends and skewed distributions that can compromise model accuracy if left unaddressed.

One key transformation is the encoding of categorical variables into numerical formats, enabling machine learning algorithms, which typically require numerical input, to process qualitative information. For instance, one-hot encoding was applied to variables like reservoir formations to represent each category as a binary vector, preserving class distinction without imposing ordinal relationships. Binary encoding was used for simpler categorical attributes like fluid types (e.g., oil, gas, water), reducing dimensionality while maintaining interpretability.

Additionally, logarithmic and power transformations were employed to normalize highly skewed data distributions. These transformations help stabilize variance, reduce the influence of outliers, and linearize exponential trends, making the data more compatible with algorithms that assume normality or linear relationships. For example, oil production rates were transformed using logarithmic scaling to reduce positive skewness, and fluid

flow patterns were numerically encoded (e.g., as values 1 through 6) to enable classification model training.

These transformation techniques not only improve model robustness and training efficiency but also enhance the interpretability and generalizability of the resulting predictions, critical considerations for real-world applications in flow pattern recognition and hydrocarbon production forecasting.

Proper data partitioning is essential for developing machine learning models that generalize well to new, unseen data, particularly in petroleum engineering, where reservoir heterogeneity, variable operational conditions, and limited sample sizes can easily lead to overfitting. To ensure robust evaluation of model performance, this study employed Repeated K-Fold Cross-Validation, a technique that balances variance and bias more effectively than a single random split.

In standard K-Fold cross-validation, the dataset is divided into K equally sized folds. Each fold is used once as the validation set, while the remaining folds form the training set. This process is repeated K times, and the final performance metric is calculated as the average of all runs (Kuhn and Johnson, 2016). However, a single K-Fold run can still be sensitive to how the data was initially shuffled or divided.

To address this, the study applies Repeated K-Fold, where the entire K-Fold process is repeated multiple times with different random shuffles of the data (Pedregosa et al., 2011). This repetition reduces the variability introduced by random splits and offers a more reliable estimate of model performance, which is particularly valuable in this study where well counts may be limited and class imbalance (e.g., flow regimes) may affect

learning. Repeated K-Fold also helps detect overfitting and provides a clearer picture of how well the model will perform in real-world scenarios.

To further enhance model development, the dataset was partitioned into training, testing, and (optionally) validation sets:

- 80% Training Set – Used to train the model and learn feature relationships.
- 20% Testing Set – Held out entirely for final performance evaluation.
- Optional 10% Validation Set – Used for hyperparameter tuning and model refinement, without leaking information from the test set.

This structured approach ensures that each phase of model development remains unbiased and that model predictions reflect generalization to new wells, not memorization of the training data. For example, in a dataset containing 415 wells (195 Wolfcamp and 220 Montney wells), the data split would result in:

- Training Set: 332 wells
- Testing Set: 83 wells
- Validation Set (if used): 41 wells

This partitioning preserves the underlying class and feature distributions while minimizing the risk of data leakage. In the context of hydrocarbon production forecasting, where unseen wells may have different stimulation treatments or geological conditions, proper partitioning is vital for building models that are both accurate and deployable in the field.

Overall, rigorous partitioning and validation methods ensure that the machine learning models in this study are developed, tested, and evaluated on fair and realistic data

scenarios. This directly supports the study's objective of producing physically consistent, generalizable, and trustworthy ML models for reservoir forecasting.

2.1.7 Model Evaluation Metrics

Evaluation metrics are essential for measuring the performance of machine learning models, allowing for quantitative assessment of how well a model generalizes to new data. In this study, both regression and classification tasks are addressed, requiring appropriate metrics for each. The chosen metrics not only measure model accuracy but also provide insight into error magnitude, variance explained, and the reliability of classification under imbalanced conditions.

For regression tasks such as cumulative production forecasting, several widely accepted metrics are used. R-squared (R^2), or the coefficient of determination, measures the proportion of variance in the target variable that is explained by the model. It ranges from 0 to 1, with higher values indicating a better fit. An R^2 of 1.0 implies perfect prediction, while a value near 0 implies the model explains very little of the variability (Kuhn and Johnson, 2016). This metric is particularly useful for understanding how well the model captures trends in reservoir behavior and completion effects.

Root Mean Squared Error ($RMSE$) quantifies the average magnitude of prediction errors, giving higher weight to larger errors by squaring the residuals. It is the square root of Mean Squared Error and maintains the same unit as the target variable. $RMSE$ is especially important in engineering applications, where large deviations from actual values (e.g., production rates) can result in costly operational decisions (Willmott and Matsuura, 2005). A lower $RMSE$ indicates better model performance.

Mean Squared Error is similar to $RMSE$ but not normalized. It penalizes larger errors more severely, making it sensitive to outliers. While it lacks intuitive interpretability due to squared units, it is valuable when detecting whether the model occasionally makes large errors—which could indicate instability in production predictions (Kuhn and Johnson, 2016).

Mean Absolute Error calculates the average absolute difference between predicted and actual values. Unlike $RMSE$, it treats all errors equally, making it less sensitive to outliers. This metric is especially useful when consistent predictive accuracy is prioritized over penalizing rare but large deviations. MAE retains the same unit as the predicted variable, offering a more intuitive sense of average prediction error (Willmott and Matsuura, 2005).

Together, R^2, $RMSE$, MSE, and MAE provide a comprehensive view of a model's performance: R^2 shows how much of the trend is captured, $RMSE$ and MSE reveal how significant the prediction errors are, and MAE reflects the model's average deviation across all predictions.

For classification tasks such as flow pattern recognition, metrics must assess not only accuracy but also how well the model performs across multiple classes and avoids bias toward majority categories. Accuracy measures the proportion of correct predictions out of all predictions. While easy to interpret and commonly used, accuracy alone can be misleading in imbalanced datasets, common in flow pattern classification where certain regimes (e.g., slug, annular) may be underrepresented. A model can achieve high accuracy by simply predicting the majority class, without truly learning minority patterns (Kuhn and Johnson, 2016).

Cohen's Kappa (κ) adjusts for agreement occurring by chance, providing a more meaningful evaluation than accuracy alone. It is especially useful when class imbalance exists. Kappa values range from -1 to 1, with 1 indicating perfect agreement, 0 meaning agreement by chance, and negative values implying worse-than-random performance (Cohen, 1960). In this study, where accurate flow pattern recognition is essential for modeling pressure profiles and equipment design, Kappa provides a clearer picture of true model skill.

For Regression (Production Forecasting): High R^2, low $RMSE$, and low MAE signify models that accurately forecast production rates and EUR across various unconventional reservoirs. These metrics are essential for validating the Machine Learning Assisted–Decline Curve Analysis (MLA–DCA) framework and ensuring reliable, field-applicable predictions.

For Classification (Flow Pattern Recognition): Metrics like Kappa and accuracy evaluate how well the model distinguishes between complex two-phase flow regimes. This is critical for pressure drop estimation, artificial lift design, and overall production system modeling.

By using multiple complementary metrics, the study ensures that model evaluation is robust, interpretable, and aligned with engineering objectives, supporting trustworthy deployment in real-world petroleum engineering workflows. Mathematical representations of these metrics are provided in **Appendix I.**

Chapter 3: Machine Learning Two-Phase Fluid Flow Pattern Identification

3.1 Research Motivation

Integrating dimensional analysis with machine learning is a major step forward in engineering and data science. Machine learning is powerful for processing large datasets and finding patterns, but it can struggle in systems governed by physical laws. Two key challenges are interpretability and generalizability.

Interpretability means understanding how a model makes its predictions. Many machine learning models act like "black boxes", making it hard to explain their decisions. This can be a problem in engineering, where results need to align with known physical laws. Dimensional analysis helps by using dimensionless groups, simple combinations of variables that represent physical relationships. These groups make the models easier to understand and their predictions more transparent.

Generalizability is about how well a model works on new data. Machine learning models often perform well on the data they’re trained on but fail in new or different scenarios. This is a big issue in engineering, where conditions like fluid properties or pipe geometry can vary widely. Dimensional analysis addresses this by creating predictors based on fundamental physics. These predictors apply across different systems, helping models perform better in real-world situations.

The flow pattern of a multiphase system refers to the spatial distribution of liquid and gas phases within a conduit as they flow simultaneously. Determining these patterns is a fundamental challenge in two-phase flow analysis and plays a critical role in optimizing hydrocarbon production and transportation. An accurate flow pattern prediction model

supports petroleum engineers in making data-driven decisions, improving production rates, and minimizing downtime.

The model presented in this work integrates machine learning with dimensional analysis to address these challenges. By using machine learning techniques trained on dimensionless features, the model achieves higher accuracy and adaptability across different operating conditions. This combination ensures that the model is not only grounded in physical principles but also applicable to a wide range of practical scenarios in industries reliant on efficient and safe gas-liquid flow systems. By combining dimensional analysis with machine learning, this work creates models that are both easier to understand and more reliable. This approach bridges the gap between theory and practical application, making machine learning a more powerful tool for solving complex engineering problems.

3.2 Machine Learning Model Details

Machine learning techniques have impacted flow pattern prediction by improving accuracy and efficiency beyond traditional empirical and mechanistic methods. By leveraging a large experimental database, ML models can identify complex relationships between fluid properties, flow parameters, and flow regimes, reducing reliance on predefined theoretical assumptions. This study integrates dimensional analysis to transform raw experimental measurements into dimensionless variables, preserving the underlying physical principles while minimizing computational complexity. The use of feature selection and dimensional reduction ensures that the most relevant variables contribute to the model, enhancing generalization and interpretability. The following sections outline the experimental dataset, dimensionless variable formulation, and ML

modeling approach used to develop a robust and scalable flow pattern prediction framework.

3.2.1 Experimental Data

The flow pattern is affected by fluid properties, in-situ liquid and gas flow rates, flow conduit geometry, and mechanical properties. Laboratory data since the 1950s have been collected, and more than 8000 data points have been obtained. However, the actual flow conditions are significantly different in any laboratory setting. Therefore, several dimensionless variables were derived to characterize these data points first. Then, machine learning techniques were applied to these dimensionless variables to develop the flow pattern prediction models. Applying hydraulic fundamentals and dimensional analysis, dimensionless groups are developed to reduce the number of freedom dimensions. These dimensionless groups are simple to use for upscaling and have physical meanings. The collected data is converted from actual laboratory measurements to variations of these dimensionless variables. Machine learning models combined with dimensionless predictors significantly improve the predictive accuracy of the ML model. Until now, the best matching method for these laboratory data has been about 80%, using the most recently developed semi-analytical models. The quality of the matching is improved to more than 90% on the experimental data using machine learning techniques.

A database of 12 laboratory experiments was compiled to explore the accuracy of predicting flow patterns. The database comprises measurements of density, viscosity, temperature, pressure, velocity, surface tension, pipe diameter, angle of the pipe, and the observed flow patterns. These patterns include dispersed bubble (DB), stratified smooth

(SM), stratified wavy (SW), annular (A), intermittent (I) and bubble (B) as depicted in the traditional boundary maps below in **Figure 4**.

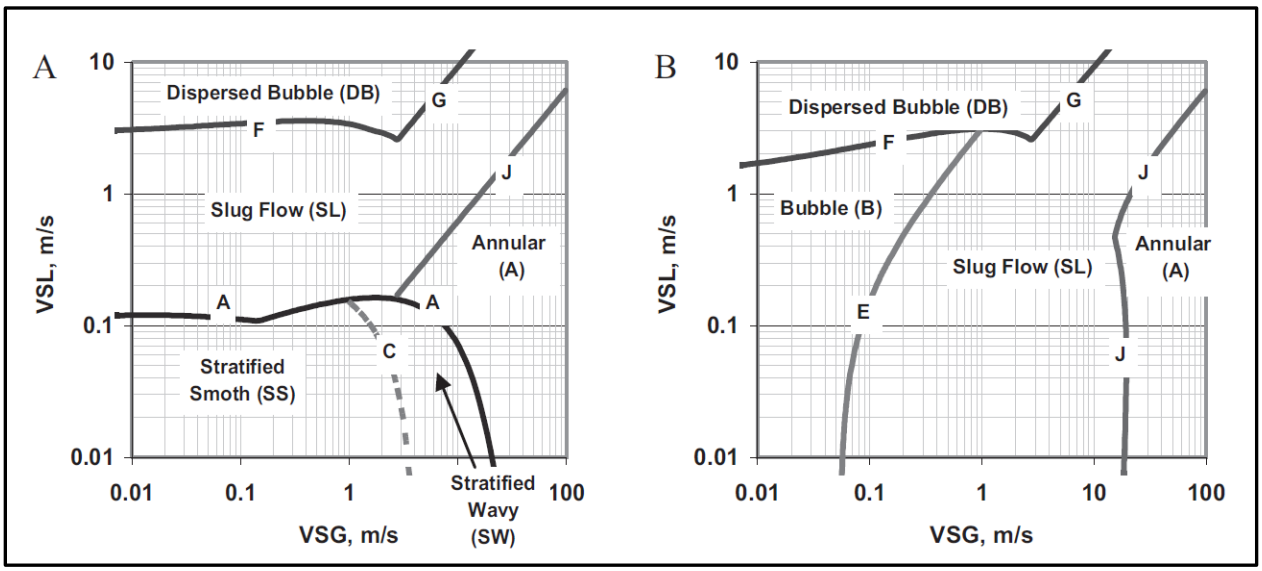

Figure 4 Transition boundaries and observed flow patterns in gas-liquid two-phase flow, including dispersed bubble, stratified smooth, stratified wavy, annular, intermittent, and bubble. These boundaries illustrate the classification of flow regimes based on experimental data (Pereya et al., 2011),

Experimental studies on gas-liquid two-phase flow in pipes have been conducted under a wide range of fluid properties, pipe geometries, and pressure conditions. These studies, compiled from various institutions, provide critical datasets for developing and validating ML-based flow pattern prediction models (Pereya et al., 2011). The dataset includes investigations covering different inclination angles, fluid viscosities, and operational pressures, ensuring a comprehensive representation of multiphase flow conditions. The table below summarizes key experimental studies, detailing the contributing authors, test conditions, and institutions involved.

Table 6. The experimental studies on gas-liquid two-phase flow in pipes, details the fluid properties, pipe geometries, pressure conditions, and institutions that contributed to the dataset. These studies provide a broad range of flow conditions for developing and validating ML-based flow pattern prediction models (Pereya et al., 2011).

Authors	Year	Variables Range	Points	Institution
Shoham	1982	Air and water; $ID = 2\ and\ 1\ in.$; from $\theta = -90^{\circ}\ to\ +90^{\circ}$	5676	TelAviv University
Lin	1982	Air and water; $ID = 25.4$ and 95.4 mm; $\theta = 0^{\circ}$	141	University Illinois
Kouba	1986	Air and kerosene; $ID = 3\ in.$; $\theta = 0^{\circ}$; $\rho_L = 814\ \frac{kg}{m^3}$; $\mu_L = 1.9\ cP$	53	University Tulsa
Kokal	1987	$ID = 25.8, 51.2, and\ 76.3\ mm$; $\theta = 0^{\circ}, \pm 1^{\circ}, \pm 5^{\circ}, and\ \pm 9^{\circ}$; $\rho_L = 858\ \frac{kg}{m^3}$; $\mu_L = 7\ cP$	1668	University Alberta
Wilkens	1997	Salt water and oil; $\theta = 0^{\circ}, 1^{\circ}, and\ 90^{\circ}$; $P = 40\ psi$	204	University Ohio
Meng	2001	Annular stratified flow transition for $\theta = 0^{\circ}, \pm 1^{\circ},\ and\ \pm 2^{\circ}$	153	University Tulsa
Manabe	2001	$P = 209.3\ and\ 464.8\ psi$; $\theta = 0^{\circ}, 1^{\circ}, and\ 90^{\circ}$	247	University Tulsa
Van Dresar and Siegwarth	2001	Nitrogen and hydrogen; $ID = 8.43\ mm$; $\theta = 1.5^{\circ}$	116	NASA
Mata et al.	2002	$ID = 2\ in.$; $\theta = 0^{\circ}$; $\mu_L = 480 cP$	80	Intevep
Abduvayt et al.	2003	Nitrogen and water; $P = 592\ and\ 2060\ kPa$; $ID = 54.9\ 106.4\ mm;\ \theta = 0^{\circ}, 1^{\circ},\ and\ 3^{\circ}$	443	Waseda University
Gokcal	2005	$ID = 50.8$; $\theta = 0^{\circ}$; $\rho_L = 889\ \frac{kg}{m^3}$; $\mu_L = 181 - 587\ cP$	183	University Tulsa
Omebere-Iyari et al.	2007	Nitrogen and water; $ID = 189\ mm$; $\theta = 90^{\circ}$; $P = 2000\ and\ 9000\ kPa$	98	SINTEF

The experimental studies outlined above highlight the diversity of flow conditions examined, including variations in pipe geometry, roughness, fluid rheology, flow rates, pressure, and temperature. These studies span a range of flow conditions, from air-water systems in horizontal pipes to nitrogen-water systems in vertical configurations, capturing a broad spectrum of multiphase flow behaviors. By incorporating a wide range of pressure levels, velocities, and pipe diameters, these datasets provide a robust foundation for training and validating ML-based models, enabling more accurate and generalized predictions of flow patterns across different operating conditions.

3.2.2 Dimensional Reduction

Dimensional analysis simplifies the modeling process by reducing complex, multidimensional data into meaningful, dimensionless groups that reflect the physical forces governing the system. This not only enhances model interpretability but also allows engineers to gain insights into the underlying physical mechanisms, such as the interplay of inertial, gravitational, and viscous forces, without losing the predictive power of machine learning (Mask et al., 2019)

Using the Buckingham Π Theorem will result in a combination of these well-known equations for this data. The Eötvös (Eo) Number is a ratio of gravitational (buoyancy) and capillary forces. The Eötvös Number is related to the Laplace number (La) where $Eo = La^{-2}$ (Awad, 2012). The Laplace number, also defined as the Suratman number (Su), is a useful property for identifying flow patterns and transition boundaries in research done by Jayawardena et al. (1997) and Balasubramanian et al. (2006). The Froude Number (Fr) is a measure of inertial and gravitational forces. Although the Froude number is considered important for free-surface flow (White, 2019), it can conceptually predict flow

patterns when the forces it represents are significant. The Reynolds (Re) number is a ratio of inertial forces to viscous forces and often used to differentiate between laminar and turbulent flow (Awad, 2012). Furthermore, additional equations can be derived from a given set of Π Groups. For example, several possible dimensionless variable outcomes are shown in **Table 7**.

Table 7. Dimensionless Numbers: Definitions and Physical Interpretations

Parameter	Definition	Qualitative Ratio
Archimedes Number	$A_r = \frac{\rho_l \Delta\rho g d^3}{\mu_l^2}$	Gravitational to Viscous
Atwood Ratio	$A_t = \frac{\rho_l - \rho_g}{\rho_l + \rho_g}$	Density differences
Bond Number	$B_o = \frac{g d^2 \Delta\rho}{4\sigma}$	Gravitational (buoyancy) to Capillary
Eötvös Number	$Eo = \frac{g d^2 \Delta\rho}{\sigma}$	Gravitational (buoyancy) to Capillary
Froude Number	$Fr = \frac{\rho v^2}{\rho g d} = \frac{v}{gd}$	Inertial to Gravitational
Laplace Number	$La = Su = \sqrt{\frac{\sigma}{g\Delta\rho}}$	Capillary to Gravitational
Modified Froude Number	$mFr = \frac{\Delta\rho g d}{\rho v^2}$	Inertial to Gravitational
Reynolds Number	$Re = \frac{\rho v d}{\mu}$	Inertial to Viscous
Richardson Number	$Ri = \frac{\Delta\rho g d}{\rho v^2}$	Buoyancy to Inertial
Suratman number	$La = Su = \frac{\rho_l \sigma d}{\mu_l^2}$	Capillary to Gravitational
Weber Number	$We = \frac{\rho v^2 d}{\sigma}$	Inertial to Interfacial

To upscale laboratory experiments to real application, one usually must apply upscaling techniques based on the similarity theorem. To identify flow patterns for different flow conditions using machine learning algorithms, rather than applying the algorithms directly

on measured attributes, it is applied to dimensionless scalable parameters. This practice has the following advantages. First, dimensionless parameters to be upscaled are usually dimensionless combinations of measured parameters. Therefore, the number of upscaling parameters is usually much less than the original measured attributes, which results in a smaller matrix and more efficient manipulation for the algorithms. Secondly, the upscaled parameters are independent and comprehensive, which reduces the linearity between parameters in algorithms. Thirdly, the upscaled parameters are physically meaningful for ratios of geometry or ratios of forces. They can be easily converted back to real dimensions in the predicting model (Mask et al., 2019).

3.2.3 Machine Learning Gas-Liquid Flow Pattern Prediction Model

In this study, a machine learning classification model was developed to identify gas-liquid flow patterns in two-phase systems using tree-based algorithms. Flow pattern recognition is inherently a categorical classification task, where the goal is to predict discrete flow regimes (e.g., slug, annular, bubbly) based on input features derived from dimensionless groups.

Supervised learning was applied using multivariate classification, where input predictors (feature vectors) were mapped to known flow pattern labels. Given a feature vector, an input consisting of key descriptive attributes of the system, the model aims to accurately assign the correct flow pattern category as frequently as possible (Dalpiaz, 2018). Among several tested algorithms, tree-based classifiers, particularly boosted decision trees, yielded the highest predictive accuracy. These models divide the predictor space into segmented regions, assigning class labels based on the majority within each partition. Boosted tree methods iteratively improve performance by learning from the errors of

previous models, making them well-suited for capturing complex, nonlinear patterns in flow behavior (James et al., 2013).

Model performance was evaluated using accuracy and Cohen's kappa statistic. Accuracy provides a basic measure of the proportion of correctly classified samples, but in multiclass classification with class imbalance, a common issue in flow regime data, it can be misleading. To address this, kappa was used as a more robust metric. Kappa adjusts for chance agreement and offers a better assessment of model reliability across imbalanced classes (Kuhn and Johnson, 2016).

A confusion matrix was also generated to visualize the distribution of true vs. predicted classes, providing detailed insight into which flow regimes were most frequently misclassified. This level of evaluation is critical in petroleum engineering applications, where accurate flow regime identification informs key operational and design decisions.

By combining physics-informed features through dimensional analysis with powerful tree-based classifiers, this model achieves both interpretability and strong generalization, advancing two-phase flow classification in complex production systems.

3.3 Experimental Procedure

The procedure begins by collecting raw data relevant to the research question. Dimensional analysis is then applied to identify and reduce the number of variables, ensuring that only the most significant dimensions are retained. Next, feature extraction techniques transform the data into a more manageable form, capturing the underlying patterns while minimizing noise. This refined dataset can then train machine learning models, facilitating improved predictions and insights into the studied phenomena.

3.3.1 Transforming to Dimensionless Predictors

Through dimensional analysis, the flow pattern (∇P) is determined as a function of its independent variables:

$$f(ID, \Delta\rho g, \upsilon_s, \sigma, \mu, \rho) \qquad (4)$$

The Buckingham Π Theorem provides the derivation of the following Π Groups (see **Appendix J** for derivation):

$$\Pi_1 = \frac{gD^2\Delta\rho}{\sigma} \qquad (5)$$

$$\Pi_2 = \frac{D^2\Delta\rho g}{\upsilon_{sj}\mu_j} \qquad (6)$$

$$\Pi_3 = \frac{\Delta\rho g D}{\rho_j \upsilon_{sj}^2} \qquad (7)$$

Using the Buckingham Π Theorem, three Π groups are derived for data mining. For this paper, the Π groups and the previously noted nomenclature will be used interchangeably. Fundamentally, it is understood that buoyancy, capillary, inertial, gravitational, and viscous forces play a key role in determining flow patterns in two-phase gas-liquid flow. Therefore, the derived Π groups are determined to be valid predictors. The groups can be explored through inspectional analysis to understand the importance of using gas or liquid properties within the Π groups. A gravitational acceleration of 9.8 m/s^2 was applied to calculations that involve gravity. The angle of the pipe is incorporated into the model by both the $\sin \measuredangle_{pipe}$ and $\cos \measuredangle_{pipe}$.

3.3.2 Feature Extraction

The data does not contain any missing variables; therefore, no imputations were performed. Although the data contained outliers, none of the data was removed for this study. Examining the dimensionless groups Π_1 is seen as the inverse of the Eötvös Number (Eo) and Π_3 is seen as the inverse of the modified Froude number (mFr). Once the dimensionless variables are determined, scatter plots and histograms are created to determine relations among the predictors (independent variables).

A box-cox transformation is applied to data that did not contain non-missing or finite values that are negative. The histograms for Π_3 (mFr), which was prepared using gas and liquid and applied a box-cox transformation. The values of ${mFr_l}^{0.046}$ and ${mFr_g}^{0.033}$ were used to optimize the data mFr_g to be viewed for distribution. **Figure 5** shows that mFr_l has an even distribution of the data, while the mFr_g is skewed to the right as seen in **Figure 6**. A scatter plot of mFr_l and mFr_g with the data filtered down to air and water in a horizontal pipe displays transition boundaries forming in the data shown in **Figure 7**. The segregation of flow patterns is seen when constraining the data to individual fluid and gas such as air and water in a horizontal pipe. It was determined to use the Π_3 group as a predictor for the model. A scatter plot of mFr_l and mFr_g with data filtered to water and air, including vertical and inclined data shows that the transition boundaries of the flow patterns are less distinct for given patterns seen in **Figure 8**. A scatter plot of mFr_l and mFr_g with data for all fluids, including vertical and inclined data, shows that the transition boundaries of the flow patterns are nearly non-existent in **Figure 9**. Additional scatter plots and histograms shown in the **appendix K.**

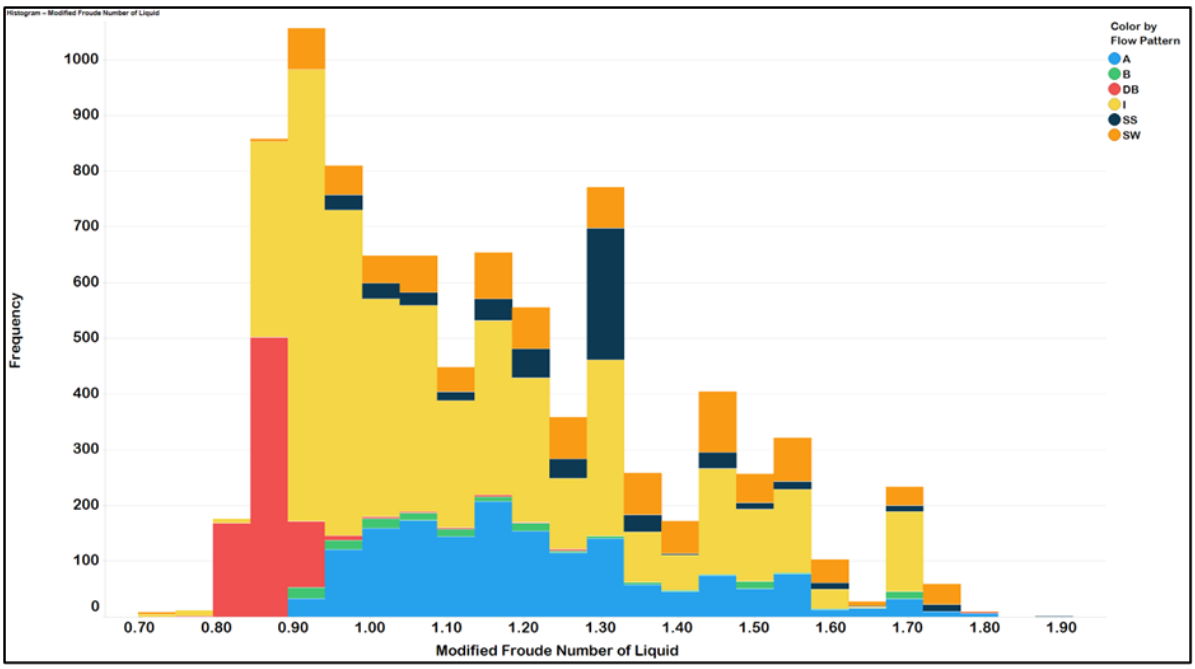

Figure 5. The even distribution of mFr_l after Box-Cox transformation, indicates its suitability as a predictor, while distinct flow pattern groupings highlight its role in displaying flow regime transition.

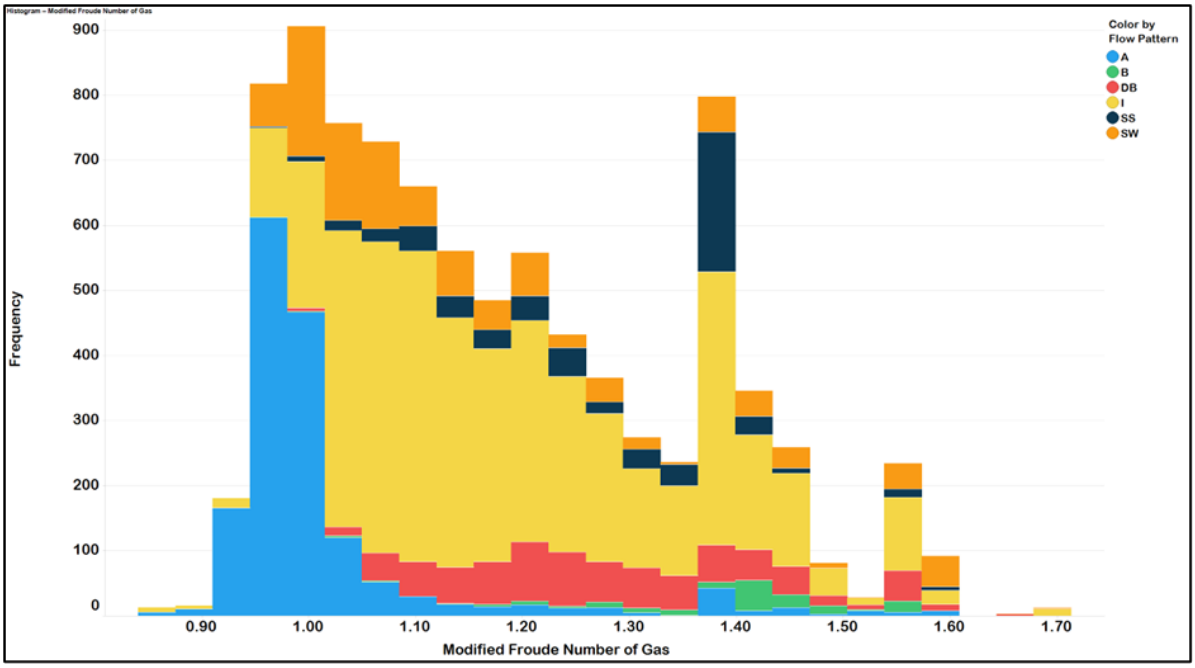

Figure 6. The slight right skew of mFr_g after Box-Cox transformation contrasts with the more uniform distribution of mFr_l, indicating differences in flow regime behavior and supporting its role as a predictive feature.

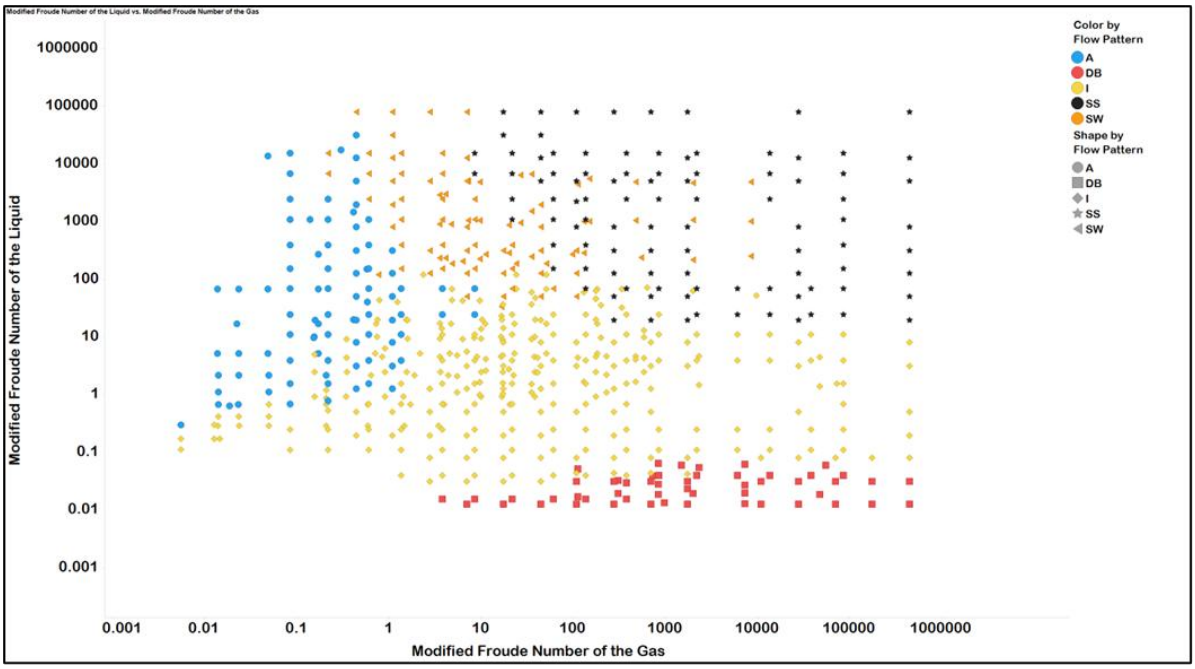

Figure 7. Distinct clustering of flow patterns gas-liquid two-phase flow for air and water in a horizontal pipe indicates transition boundaries, supporting the use of mFr_l vs. mFr_g as key features for flow regime classification.

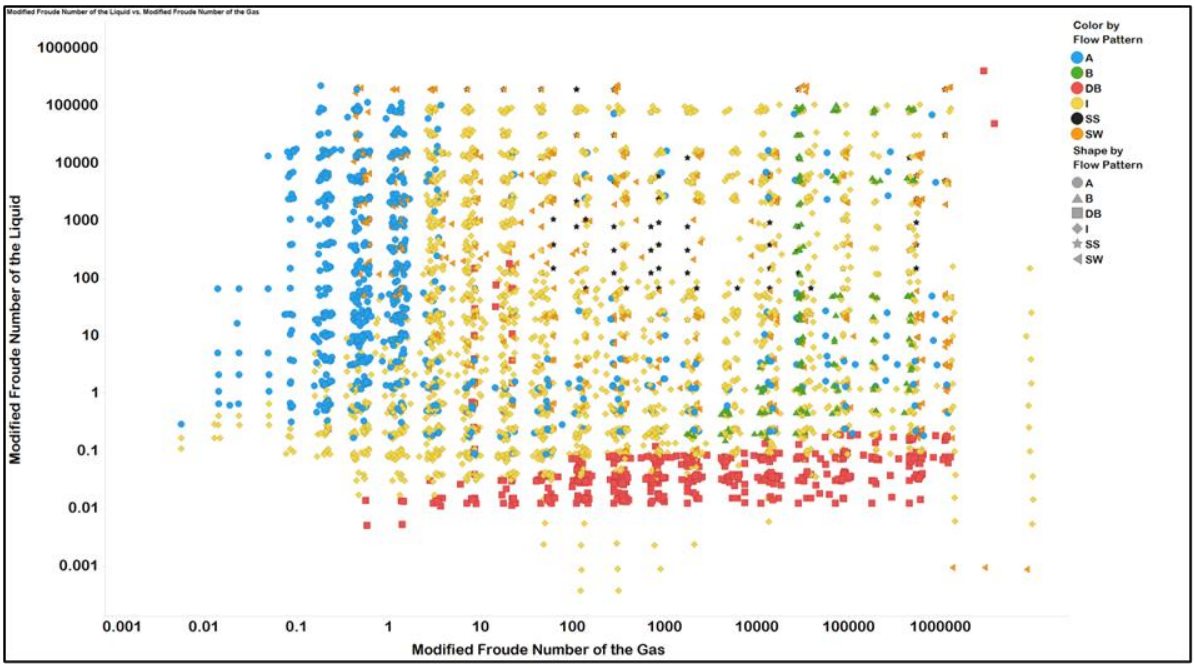

Figure 8. The transition boundaries of the mFr_l vs. mFr_g for air and water in all pipes between flow patterns become less distinct, indicating increased complexity when incorporating additional flow orientations.

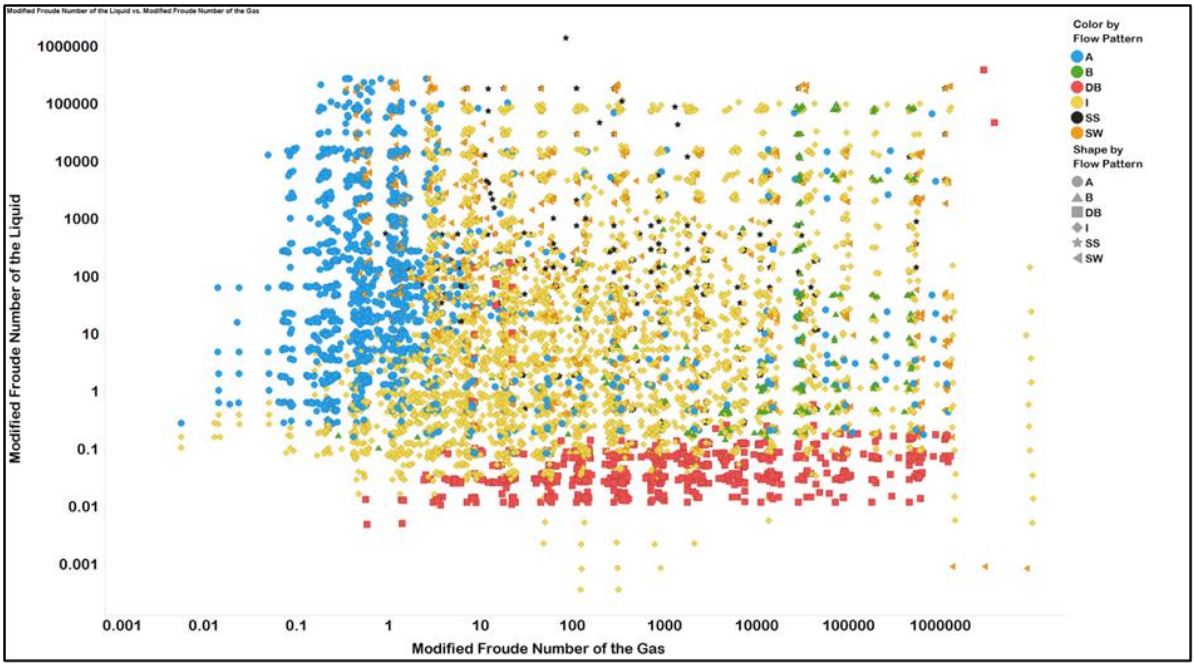

Figure 9. The transition boundaries of the mFr_l vs. mFr_g of all fluids plot shows flow patterns become less distinguishable, indicating that additional complexities arise when generalizing across different flow conditions and fluid types.

The flow patterns shift away from distinct separation when viewing the air-water at different angles. Finally, the pattern boundaries are no longer held when viewing all the liquids and gases together at varying angles. Interestingly, while the boundaries are not distinct, the annular and dispersed bubble patterns are consistent for the regions where they fall. It is observed that the angle of the pipe influences the flow pattern boundaries when changed. Therefore, the impact of the angle is explored by incorporating it into the predictors for the ML model. The $\sin \measuredangle_{pipe}$ and $\cos \measuredangle_{pipe}$ are multiplied by the Π_1 group for inspectional analysis. While exploring using $\cos \measuredangle_{pipe}$ or $\sin \measuredangle_{pipe}$ individually, the model yielded results in the 80% range. The logic behind using both the $\sin \measuredangle_{pipe}$ and $\cos \measuredangle_{pipe}$ multiplied by E_o was determined by a simple observation. Given that the $\sin \measuredangle_{pipe}$ of zero

degrees is zero and the $\cos \measuredangle_{pipe}$ of ninety degrees is zero, it is determined that two predictors incorporating the $\sin * \measuredangle_{pipe}$ and $\cos * \measuredangle_{pipe}$ are needed for the model to differentiate between horizontal and vertical flow. By incorporating $\Pi_1 * \sin * \measuredangle_{pipe}$, and $\Pi_1 \cos * \measuredangle_{pipe}$, the model achieved an accuracy in the 90% range. The model gains considerable confidence and accuracy when using both predictors for comparison.

The model with the least number of predictors will be more efficient and less prone to overfitting. The predictors chosen for the optimal model are mFr_l, mFr_g, $Eo * \sin \measuredangle_{pipe}$, and $Eo * \cos \measuredangle_{pipe}$.

3.3.3 Experimental Models Design

The optimal model will consist of 4 predictors and 1 target. The data set is split between the train predictors and targets. The predictors include the mFr_l, mFr_g , $Eo * \sin \measuredangle_{pipe}$, and $Eo * \cos \measuredangle_{pipe}$. The target is the flow pattern. The train control uses repeated k-fold cross-validation. Individual grid models are built for each package to allow for tuning of the hyper-parameters. While Π_2 is not included in the optimal model as a predictor; the model was run using 6 predictors using Π_2 of the liquid and gas as the 5th and 6th predictors which yielded a negligible increase in accuracy.

Several tree-based machine learning algorithms are investigated, specifically Random Forest, Extreme Gradient Boosting (XGB), and AdaBoost. Each model undergoes hyperparameter tuning to optimize performance, utilizing techniques such as grid search or Bayesian optimization to identify the best parameter configurations. To ensure the robustness and generalizability of the results, repeated k-fold cross-validation is employed, which allows for multiple iterations of data partitioning and model training. The

model focuses on minimizing variance and providing a reliable estimate of model performance across different subsets of the dataset. This comprehensive approach enhances the credibility of the comparative analysis of these algorithms.

3.4 Results and Discussion

The results demonstrate that combining dimensionless variables with machine learning effectively predicts flow patterns, often surpassing the accuracy of prior semi-analytical models. The confusion matrix in **Table 8** presents the performance of the XGB tree model using 5-fold, 10-repeated cross-validation, with correctly classified targets highlighted in green. Most flow patterns are predicted with $\geq 90\%$ accuracy, with intermittent flow being the most accurately and abundantly classified. This pattern also exhibits significant influence on the prediction of other flow types, clustering around all flow patterns. In contrast, bubble flow shows the lowest accuracy and data representation.

Table 8. Confusion matrix for the XGB tree model using 5-fold, 10-repeated CV showing classification performance. The model achieves $\geq 90\%$ accuracy for most patterns, with intermittent flow having the highest accuracy and bubble flow the lowest.

Prediction	Dispersed Bubble	Stratified Smooth	Stratified Wavy	Annular	Intermittent	Bubble Flow
Dispersed Bubble	7,433	0	28	3	541	2
Stratified Smooth	0	5,210	80	67	251	0
Stratified Wavy	19	141	9,750	387	182	0
Annular	0	82	606	14,606	660	1
Intermittent	708	387	436	1,247	44,040	256
Bubble Flow	0	0	0	0	66	1,271
Prediction %	91%	90%	89%	90%	96%	83%

The outcomes of the tree-based models are seen in **Table 9**, and the models achieved up to 93% accuracy with a Kappa of 0.90. While these methods need further improvement and validation, the implications are substantial in creating powerful flow pattern prediction tools. The XGB model yielded the best results. The accuracy of the model was 93%, with a Kappa of 0.90. The Random Forest model achieved an accuracy of 92% with a Kappa of 0.89, and the Bagged and Boosted model achieved an accuracy of 92% with a Kappa of 0.88, further validating the model given the high accuracy and kappa.

Comparative analysis reveals that the model incorporating six predictors achieves an accuracy of 93% and a Kappa statistic of 0.91 when utilizing the XGB tree model. While these predicted values demonstrate a modest improvement over the baseline optimal model, this enhancement may compromise the overall efficiency of the optimal configuration. Furthermore, the practical significance of this improvement remains uncertain, particularly when considering the scalability of the model to larger datasets or real-world applications. Careful evaluation of the trade-offs between accuracy and computational efficiency is essential in determining the viability of this approach for broader deployment.

Table 9. Performance metrics of tree-based models using 5-fold and 10-fold, 10-repeated cross-validation. All models demonstrated over 90% accuracy using only four dimensionless predictors.

Classification	**5-fold 10 repeats**		**10-fold 10 repeats**	
Model	**Accuracy**	**Kappa**	**Accuracy**	**Kappa**
Random Forest	92%	0.88	92%	0.89
Boosted Bagging	91%	0.87	92%	0.88
Extreme Gradient Boosting	93%	0.90	93%	0.90

Given the results of the analysis, the tree-based models are recommended utilizing the four dimensionless predictors; mFr_l, mFr_g, $Eo * \sin \measuredangle_{pipe}$, and $Eo * \cos \measuredangle_{pipe}$. The model has the least predictors based on fundamental fluid flow equations derived from dimensional analysis. Overall, the project was successful in exceeding 90% accuracy with a high degree of confidence matching the experimental observations. By obtaining this level of accuracy with multiple models, the results yield a higher level of confidence in the methodology. Using only four dimensionless predictors, the model is robust enough to apply to many two-phase liquid gas applications and scalable to field level applications. The results show that combining machine learning and dimensional analysis creates a powerful tool backed in science and logic to predict fluid flow patterns.

Chapter 4: Machine Learning Enhanced Hydrocarbon Production Forecasting

4.1 Research Motivation

In modern petroleum engineering, rapidly evolving challenges in unconventional reservoirs, such as heterogeneous rock properties, diverse completion designs, and, shifting operational protocols, demand forecasting tools that can adapt quickly and reliably. While machine learning provides a powerful framework for capturing nonlinear relationships and managing large datasets, traditional implementations can be cumbersome, computationally intensive, and difficult to customize.

Many reservoir settings lack extensive data, making traditional simulation approaches or purely data-driven models insufficiently robust. By incorporating dimensional analysis, engineers can reduce model complexity while preserving critical physical insights, enhancing both interpretability and scalability. When combined with transfer learning, this approach enables the mapping of scalar EUR estimates to full cumulative production profiles, enhancing prediction accuracy across time-series outputs. Additionally, integrating empirical decline curve analysis offers reliable late-time forecasts and a fallback method when ML-based predictions become unreliable outside their training domain. This synergy of engineering expertise and data science ensures a workflow that optimally balances computational efficiency with physical realism, particularly vital in heterogeneous and rapidly evolving reservoir environments.

The goal is to streamline the process of building and executing ML models for production forecasting through a structured, step-by-step methodology that integrates Dimensional Analysis, Transfer Learning, and empirical decline curve models. This combined

approach aims to reduce computational costs and data requirements, while maintaining model interpretability and robust long-term predictions. Crucially, it is not about identifying a single “best” model or definitive set of hyperparameters. Rather, the intent is to equip petroleum engineers with a practical workflow that can be adapted to their unique reservoirs, data availability, and expertise, ultimately enhancing accuracy, reliability, efficiency, and transferability of production forecasts.

4.2 Methodology

This study presents a framework to predict hydrocarbon production for MFHWs by improving the efficiency of the ML algorithm and expediting the training process by combining dimensional analysis, transfer learning, and decline curve analysis methods. The proposed method uses dimensional analysis to develop input predictors with physical meaning, create a more generalized model by applying transfer learning, and extend forecasts into late time using empirical modeling. The steps are presented as follows:

1. **Collect and Preprocess Data:**
    a. Gather raw data from reliable sources as determined by the operator.
    b. Handle missing values by employing techniques such as mean imputation or interpolation.
    c. Normalize or scale targets to ensure uniformity in the data.
2. **Define Dimensionless Predictors Utilizing Dimensional Analysis:**
    a. Identify relevant variables related to completions and or reservoir characteristics.

b. Apply Buckingham Π Theorem to derive dimensionless groups from these variables.

c. Validate dimensionless groups to ensure they capture essential physical relationships.

3. Deploy Scalar Machine Learning Model for EUR Prediction:

a. Incorporate dimensionless predictors into a scalar model.

b. Utilize machine learning techniques to identify regression algorithms such as linear regression, random forest, or neural networks.

c. Apply techniques such as grid search and repeated K-fold cross-validation to fine-tune and validate an optimal regression algorithm and corresponding hyperparameters.

d. Validate the EUR predictions using statistical evaluation metrics such as R-squared or Root Mean Square Error.

4. Apply Transfer Learning – Scalar to Vector Machine Learning Model:

a. Transfer the knowledge (regression algorithm and hyperparameters) from the trained scalar model to initialize the vector model.

b. Incorporate dimensionless predictors from the scalar model into the vector model, noting that the dimensionless predictor that contains time is the dynamic predictor. This dynamic predictor encapsulates temporal attributes, thus distinguishing it as a key component within the ensemble of dimensionless predictors.

c. Validate the modified dimensionless predictors to maintain consistency with the underlying physical processes.

d. Adapt the vector model to accept the input data with multiple predictors, retraining the learned representation from the scalar model.

5. **Deploy Vector Machine Learning Model for Cumulative Production Over Time Predictions:**

    a. Identify the location and properties (reservoir and or completions) of the proposed undrilled well to establish the input parameters for prediction.

    b. Utilize the newly developed vector model to predict cumulative production over time for the identified undrilled well with established production-type curves, employing traditional methods such as DCA to assess the goodness of fit and validate the predictive accuracy of the vector model.

    c. Validate the accuracy of the vector ML model by training the selected regression algorithm on different datasets.

    d. Observe the generalization of the vector ML model by training the selected regression algorithm on different datasets.

6. **Validate Transfer Learning – Vector Model**

    a. Conduct sensitivity analysis to validate knowledge transfer in vector ML-based production forecasting, confirming the model's ability to effective use key features.

b. Identify dynamic input predictor which is essential for accurate forecasting due to its influence on production curves.

c. Assess sensitivity outcomes from vector model as they should align with the scalar features of importance ranking to validate knowledge transfer.

7. Implement Machine Learning Assisted – Decline Curve Analysis:

a. Use a machine learning model to predict cumulative production over time where sufficient data is available for training.

b. Use the decline curve analysis model to predict cumulative production over time where data is unavailable to train the ML model.

c. Concatenate the ML predictions generated for the initial 36 months with the DCA prediction for the subsequent 36 months noting the initial and subsequent times are determined by the operator and data availability.

d. Validate the MLA-DCA method by comparing the forecasted and analogous production-type curves.

Following these steps, the operator can systematically preprocess the data, derive dimensionless predictors, build a regression model, implement transfer learning, incorporate temporal aspects, and generate long-term production forecasts using the MLA-DCA method. The key constraints of the MLA-DCA method include its reliance on dimensional analysis, which depends on the accuracy and completeness of input data, and the need for careful selection of source data in transfer learning to avoid negative transfer effects. Additionally, the method's use of empirical DCA models may limit its

ability to capture complex, nonlinear production behaviors in unconventional reservoirs if sufficient training data is lacking for the ML model.

The MLA-DCA method offers several advantages over pure machine learning and traditional DCA methods. It integrates dimensional analysis with ML models, enhancing interpretability and providing dimensionless predictors with clear physical meaning. Transfer learning allows the model to generalize effectively in data-scarce regions, reducing the need for large datasets. Transfer learning also identifies the regression algorithm and corresponding hyperparameters expediting the development of the ML model. Additionally, MLA-DCA balances early-time and late-time forecasting, using ML for short-term predictions and DCA for long-term, while simplifying input parameters to reduce computational complexity. The MLA-DCA method provides a unique combination of the strengths of both pure ML and traditional DCA techniques. It enhances interpretability, reduces computational complexity, and improves long-term forecasting through an integrated approach.

4.3 Experimental Data

The Wolfcamp and Montney formations are selected as case studies because they represent tight oil and shale gas resources. Data for these formations was from literature and Enverus (Nieto et al., 2009; Ojha et al., 2017; Vocke et al., 2018; Akai and Wood, 2014; Vaisblat et al., 2021; Woo et al., 2021; Vaisblat et al., 2022). The collected data set includes parameters such as oil production, gas production, time, oil API, fluid specific gravity, reservoir formation name, latitude, longitude, total vertical depth (TVD), stage spacing (S), average hydraulic fracture treatment pressure (P_T), hydraulic fracture treatment rate (Q_T), clusters per 1000 ft, shots per 1000 ft, lateral length, pounds of

proppant pumped, barrels of fluid pumped (V_f), pounds of friction reducer, mudweight, and depth from the top formation (DFT). Proppant volume (V_p) and treatment fluid viscosity (μ_T) (Sanders et al.) are derived from the data. The lateral length is constrained to wells greater than 2000 ft. long and less than 15,000 feet long to represent the multi-fractured horizontal wells studied. The completion parameter statistics compiled from completions reports are displayed in **Table 10**. The training data distribution shows the range of completions best suited to forecast production for a new well. Ensuring that the ML model follows a distribution similar to future test samples is crucial for the machine learning model's reliable learning process (Mitchell, 1997). In addition, to mitigate overfitting and facilitate accurate predictions within the same training distribution, a minimum production period of three years is applied.

Table 10. The statistical values for hydraulic fracturing treatment parameters used in the dataset help define the distribution of well characteristics used for training the machine learning model, ensuring that production forecasts align with real-world completion conditions.

Summary Value	Pressure (psi)	Rate ($\frac{bbl}{min}$)	Proppant Volume (f^3)	Fluid Volume (bbl)	Fluid Viscosity (cp)	Stage Spacing (ft.)
Variance	2.44×10^6	186	3.31×10^7	1.75×10^8	3.64	2.48×10^4
Min	3954	37	2600	6280	4.0	56
Mean	7161	67	1.15×10^4	2.58×10^4	6.04	323
Max	1.10×10^4	101	4.78×10^4	9.05×10^4	20.0	925
Standard Deviation	1562	13.3	5757	1322	1.91	157

The raw dataset comprises approximately 195 Wolfcamp wells and 220 Montney wells. The statistical overview of reservoir characteristics is presented comprehensively in **Table 11.** Additionally, parameters such as slurry volume (V_s), slurry density (ρ_s) and gravity (g) are scrutinized.

Table 11. The statistical overview of key reservoir characteristics in the dataset highlights the variability in reservoir properties. Understanding these distributions is essential for modeling production behavior and optimizing machine learning predictions.

Summary Value	Reservoir Pressure (psi)	Fluid Density ($\frac{lbm}{ft^3}$)	Fluid Viscosity (cp)	Permeability (nD)	Porosity	DFT (ft.)
Variance	5.3×10^5	602	0.64	2.20×10^8	3.1×10^{-5}	56662
Min	2805	0.10	1.2×10^{-2}	278	0.032	31
Mean	4142	23	0.70	1.34×10^4	0.045	466
Max	6675	51	3.45	3.0×10^4	0.058	1469
Standard Deviation	728	24.53	0.798	1482	0.006	238

4.3.1 Data Preprocessing

Addressing the inherent complexities and data quality deficiencies necessitates rigorous data preprocessing for valid data analyses (Fan et al., 2021). During this preprocessing phase, the operator identifies inaccurate, incorrect, or incomplete data entries. This study employs imputation from like-kind wells where well data is present to address missing data. Normalization is applied to production time, commencing at zero, while production values are normalized to the fluid volume per 1000 ft of completed lateral length, resulting in normalized oil and gas volumes. After filtering the data set down, the missing data is addressed. Individual well porosity and permeability is unavailable, but analogous data is retrieved from case studies. For the Montney wells, the permeability ranges from 10 to 30

microDarcies orders of magnitude higher than “true shale” measured in nanoDarcies, and the porosity ranges from 4 to 6 percent (Woo et al., 2021; Vock et al., 2018; Akai and Wood, 2014; Nieto et al., 2009). The average porosity and permeability for respective formations are then input randomly into the data set using Monte Carlo simulation. While the Wolcamp data is very complete, the Montney data has some missing values for DFT, mud weight, SG_{gas}, and pounds of friction reducer. The mean values are acquired from like kind wells and input for the missing data. Production time is normalized to start at zero, and production is normalized to volume of oil or gas per 1000 ft of completed lateral length. Further calculations are done to transform the variables into the completion and reservoir variables of interest. The volume of the slurry is calculated by converting the pounds of proppant to volumes using the density of the sand and adding it with the volume of total fluid pumped then converted to volume per 1000 ft lateral length. The slurry density is derived from the slurry volume and multiplied by gravity to account for gravity affects.

The viscosity values obtained by the correlation are consistent with the expected values of a light crude oil in the Wolfcamp formation. Stage spacing, clusters, and shots are also in terms of per 1000 feet of completed lateral length. The data is examined for outliers. The raw data is examined for outliers based on comparable lateral lengths, fluid properties, clusters, shots, formation, and production. The mathematical transformations of these properties are in **Appendix L.**

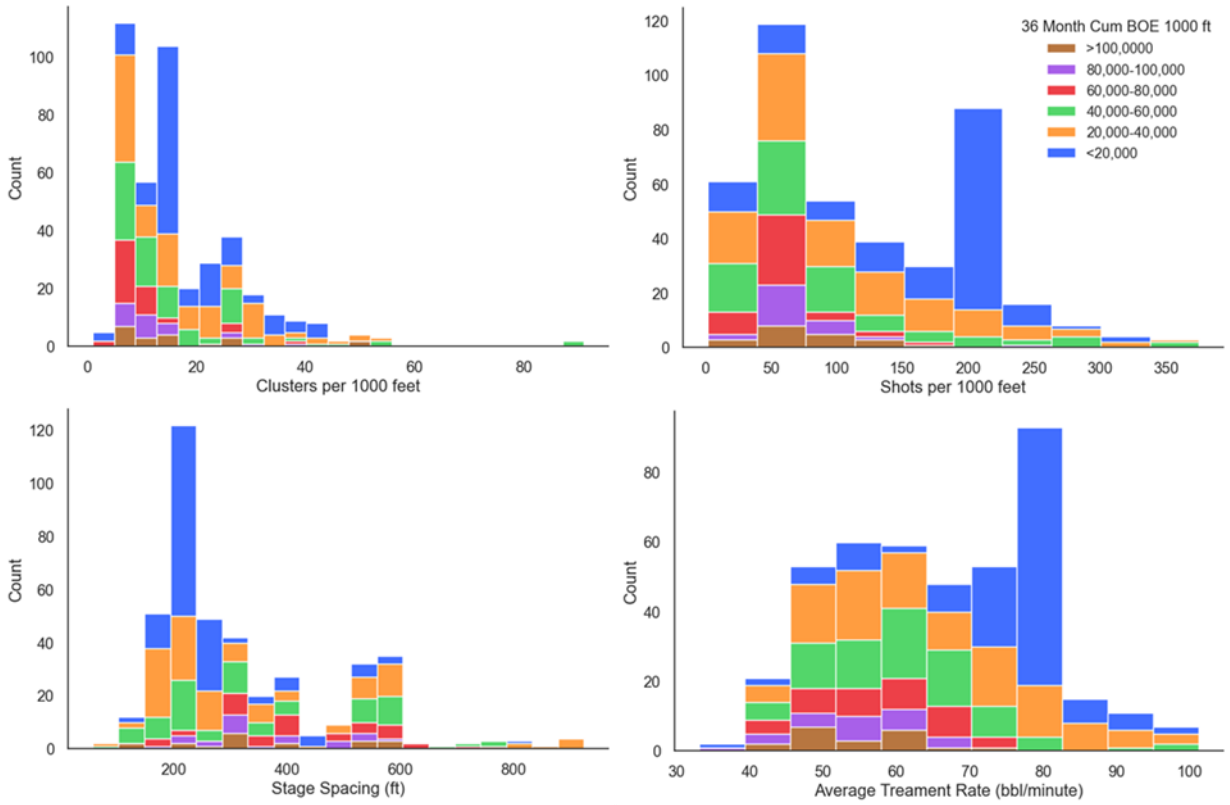

Figure 10. Distributions of key completion parameters how some variables present skewed distributions while others appear more normally distributed. Higher-productivity wells cluster in distinct parameter ranges, highlighting potential “sweet spots” and offering insights for optimizing completion designs.

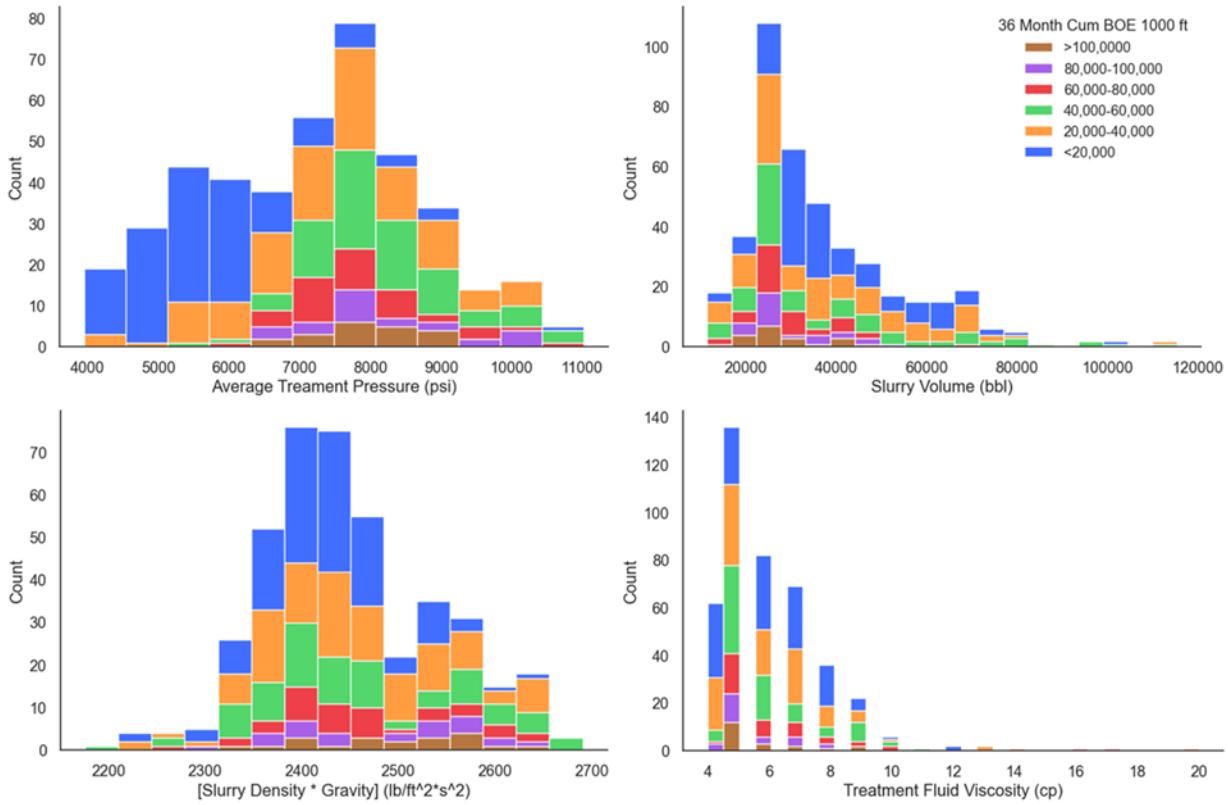

Figure 11. Histograms of key treatment parameters revealing how different operating ranges correlate with production outcomes. These distributions show that certain properties are more often associated with higher productivity, offering insights that can guide well design optimization.

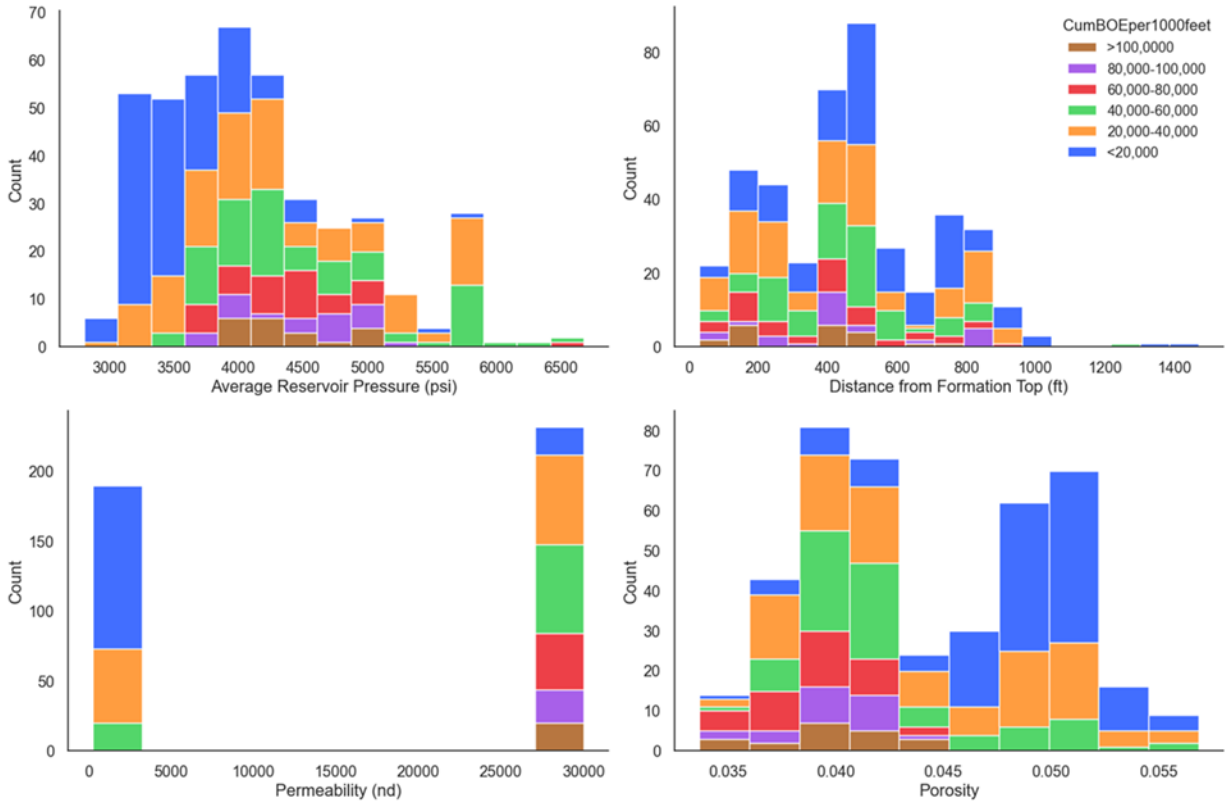

Figure 12. Distribution of reservoir properties highlights their impact on production performance. The trends indicate that higher-producing wells are associated with specific ranges of these key reservoir characteristics. The permeability distribution exhibits a bimodal pattern, reflecting contributions from two distinct reservoirs with differing permeability characteristics.

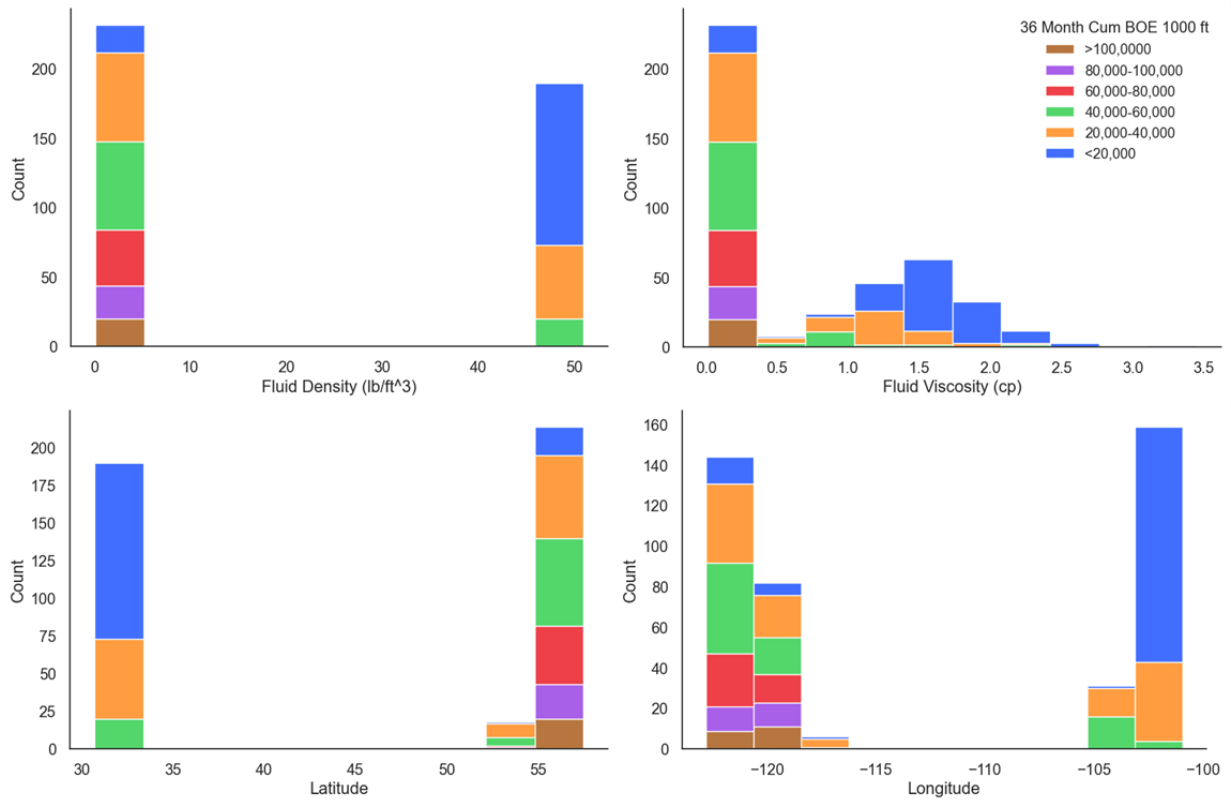

Figure 13. The histogram highlight relationship between key reservoir and fluid properties with production performance. The bimodal distribution in fluid density and viscosity reflects differences between oil and gas formations. The latitude and longitude distributions indicate that the analyzed wells are in distinct geological regions further emphasizing formation-dependent variations in production characteristics.

4.3.2 Dimensional Analysis for Feature Reduction

In this study, dimensional analysis is not applied to the geospatial data, specifically latitude and longitude, as these variables are retained to account for the spatial variability between wells. The Buckingham Π Theorem is applied separately to the completion and reservoir data. By examining these datasets independently, the remaining sixteen variables are reduced to three dimensionless completions groups and three dimensionless reservoir characteristics groups. This approach allows us to maintain physical relevance of each dataset while simplifying the input parameters used in the ML

model. The application of this method is at the operator's discretion and guided by engineering knowledge, ensuring that only the most relevant variables for predicting hydrocarbon production are included. This tailored approach enhances model efficiency and interpretability without complicating the model with unnecessary data.

Buckingham Π Theorem is applied to generate dimensionless predictors. The model starts with eighteen input parameters for the selected model. Upon applying DA, the number of input parameters is reduced to eight, including latitude and longitude.

The model considers the completions as a function:

$$f(P_T, Q_T, V_s, \mu_T, \rho_s * g, S), \quad (8)$$

The model considers the reservoir performance as a function:

$$f(P_i, \rho_f, \mu_f, k, \phi, D, t), \quad (9)$$

Table 12 shows the derived dimensionless groups for production analysis. **Appendix M** provides the mathematical derivation of the reservoir groups using Buckingham Π Theorem. Porosity, clusters per foot, and shots per foot can be incorporated into dimensionless groups where Π_1 is multiplied by clusters per foot (N_c), Π_2 is multiplied by shots per foot (N_s), and Π_4 is multiplied by porosity (ϕ).

Table 12. Dimensionless parameters derived using dimensional analysis combine and simplify the predictive features.

Dimensionless groups	Physical meaning	Category
$\Pi_1 = \frac{P_T}{\sqrt[3]{V_s}\,\rho g}$	Force of the Hydraulic Stimulation	Well completion
$\Pi_2 = \frac{S}{\sqrt[3]{V_s}}$	Volume of the Hydraulic Stimulation	Well completion
$\Pi_3 = \frac{Q_t \mu_f}{V_s^{\frac{4}{3}} \rho g}$	Rate of the Hydraulic Stimulation	Well completion
$\Pi_4 = \frac{k P_i \rho}{\mu^2}$	Storage Capacity and Deliverability	Reservoir characterization
$\Pi_5 = \frac{D}{k^{\frac{1}{2}}}$	Depth within zone	Reservoir characterization
$\Pi_6 = \frac{P_i t}{\mu}$	Reservoir Fluid Flow	Flow dynamics

4.4 Transfer Learning – Scalar to Vector Machine Learning Model

Transfer learning presents a practical and efficient approach for production forecasting, especially in early-stage reservoir development. Mask et al. (2025) integrated transfer learning into their Machine Learning Assisted – Decline Curve Analysis method, demonstrating how it can improve early-stage production forecasting. The study emphasized that transfer learning facilitates hydrocarbon production predictions by leveraging knowledge from data-rich reservoirs to accelerate model training for unconventional reservoirs with sparse data. They also introduced a scalar-to-vector transfer learning approach, which enhances model generalization by transferring knowledge from simpler regression models (EUR predictions) to more complex cumulative production forecasting tasks (Mask and Wu, 2023). However, selecting

appropriate source datasets, mitigating negative transfer, and ensuring domain similarity are critical to achieving optimal forecasting accuracy.

The proposed transfer learning approach leverages the trained machine learning model from the source task (EUR prediction) and applies it to the target task (cumulative production curves). This process allows the model to generalize effectively by reusing the learned regression algorithm and hyperparameters, reducing the time and computational resources needed to train a new model from scratch (Tan et al., 2018; Cornelio et al., 2023). Transferring the scalar EUR model's regression algorithm and hyperparameters to the vector cumulative production model enables efficient and accurate production forecasting for undrilled wells in new areas.

The transfer learning process begins by training a scalar model to predict EUR based on dimensionless predictors derived through dimensional analysis. This scalar model is adapted to a vector model that predicts cumulative production curves over time. The transferred knowledge from the scalar model improves the accuracy and efficiency of the vector model, allowing it to forecast production for wells in the target field. By systematically applying transfer learning, the model can leverage data from analogous wells or formations, facilitating more reliable forecasts in early-stage development fields where well data is limited. This method effectively mitigates overfitting and enhances model generalization, making it a power tool for predicting hydrocarbon production trends in unconventional resources (Odi et al., 2021; Cornelio et al., 2023; Mask et al., 2023).

4.4.1 Static Machine Learning Model Tuning and Validation

The algorithms are optimized by tuning hyperparameters to improve performance, such as maximizing R^2 score or minimizing $RMSE$. Cross-validation serves as a crucial step in the analysis. It is a data resampling method utilized to assess the predictive model's generalization ability, as noted by Saleh et al. (2022). The primary objectives of cross-validation include detecting overfitting, evaluating accuracy, and assessing the precision of the ML model. This process must be performed for each regression algorithm.

Hyperparameter tuning optimizes the machine learning model by adjusting its hyperparameters. A grid of potential hyperparameters is defined. A random search cross-validation algorithm is then used to explore these combinations and examine their statistical evaluation metrics. The model finds the optimal configuration of hyperparameters for improved predictions resulting in a high R^2. The results are systematically organized and analyzed to ensure relevant hyperparameter values are selected for the final model's evaluation and prediction.

After the hyperparameters are defined, Repeated K-Fold Cross-Validation is performed using the selected hyperparameters. Repeated K-Fold Cross-Validation is a variant of K-Fold Cross-Validation that enhances the robustness of the model's evaluation by adding a layer of randomness (Pedregosa et al., 2011).

4.4.2 Static EUR Machine Learning Model

The parametric comparison study aims to predict 3-year oil EUR per 1000 ft lateral length using ML models, the last point on the cumulative production curve. The study compares the results using Full, DA, and PCA predictor integrated ML models to show how the

proposed method compares with other methods. The Full model uses all eighteen input parameters without dimensional reduction. The DA model uses the six Π groups, latitude, and longitude as features. The PCA model is reduced to the top eight principal components compared to the DA model, which uses eight predictors. **Table 13** summarizes the results using Repeated K-Fold (5-fold, ten repeated) CV as shown in **Appendix N** with a 20% test and 80% train split for parametric study, displaying the R^2 for test and train scores predicting EUR. The analysis revealed that Dimensional Analysis outperformed Principal Component Analysis as a dimensionality reduction technique integrated with the ML model.

Table 13. $R\text{^}2$ test and train scores for Repeated K-Fold (5-fold, 10-repeated) CV in predicting EUR indicate that the DA model outperforms the PCA model as a dimensionality reduction technique.

Regression	**DA**		**PCA**		**Full**	
Model	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**
Ordinary Least Squares	0.38	0.43	0.38	0.43	0.50	0.58
Random Forest	0.65	0.94	0.55	0.92	0.66	0.96
Randomized Decision Trees	0.66	0.87	0.57	0.84	0.66	0.88
Epsilon-Support Vector	0.41	0.49	0.41	0.49	0.54	0.76
Gradient Boosting	0.62	0.90	0.51	0.90	0.63	0.93
Recurrent Neural Network	0.46	0.75	0.43	0.75	0.17	0.95
Multilayer Perceptron Neural Network	0.47	0.75	0.43	0.75	0.17	0.96

4.4.3 User Transfer Learning

Transfer learning is employed to expedite the prediction of cumulative production curves for undrilled wells. Transfer learning is a technique where knowledge gained from solving

one problem, known as the source task, is applied to solve a related problem, the target task (Pan and Yang, 2010; Weiss et al., 2016). In the context of this study, the source task involves predicting the estimated ultimate recovery for wells in a data-rich region or formation, where ample historical production data is available. The target task is predicting cumulative production curves for undrilled wells in early-stage development fields, where data is scarce (Odi et al., 2021; Cornelio et al., 2023; Mask et al., 2023; Misra et al., 2024).

In the selected static ML model, there are eight feature inputs, including derived six Π groups, latitude, and longitude. The inclusion of geospatial data implicitly accounts for reservoir and geological data that may be otherwise unavailable. The geospatial data establishes a relationship between wells that are close to each other. The CV results for the ML model integrated with DA are shown in **Appendix N**.

While the extra-trees (randomized decision trees) algorithm yields slightly higher test scores ($R^2 \geq 0.60$) than the random forest algorithm (**Figure N1**), the train scores are significantly higher ($R^2 \geq 0.90$) for the random forest algorithm (**Figure N2**). The random forest regressor algorithm and coinciding hyperparameters in **Appendix O** are selected to forecast new cumulative production curves for undrilled wells based on their ability to predict 3-year oil EUR consistently and accurately.

The $RMSE$ represents how far, on average, the residuals are from zero (Kuhn and Johnson, 2016). **Figure 14** shows the residuals obtained from one iteration of the cross-validation process for EUR. The residual errors are the difference between the observed values and predicted values. Minimizing these errors enhances the model's quality, leading to more accurate predictions.

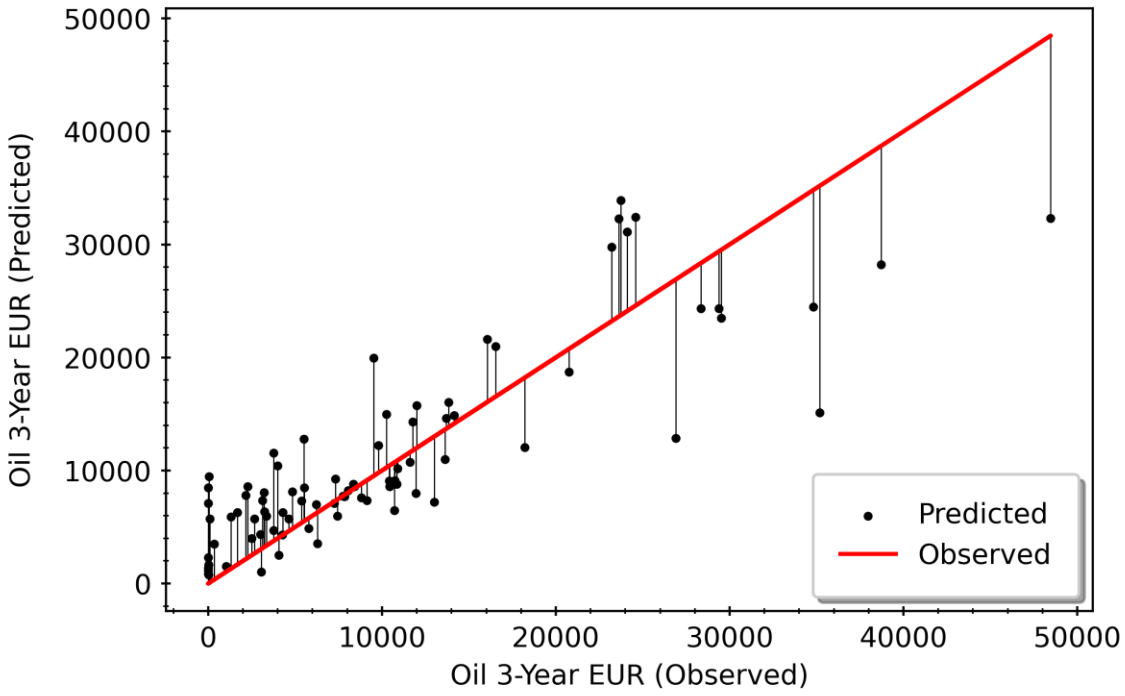

Figure 14. Presents a parity plot comparing predicted and observed EUR values, where the red line represents the ideal 1:1 correlation and the black dots denote model predictions. Deviations from the red line highlight discrepancies, offering insights into model accuracy, bias, and potential sources of errors.

Feature importance measures are employed to gain insight into how complex ML models make their predictions and to identify which features are most influential in the data generation process (Agarwal et al., 2023). By assessing the features' importance, the predictive strength of the input variables can be understood, potentially leading to a reduction in the number of predictors, thus improving the performance of the random forest regression algorithm. These importance measures are computed using the mean and standard deviation of accumulation of the impurity decrease within each decision tree (Pedregosa et al., 2011). **Figure 15** shows the predictive inputs ranked by contribution to predicting the target. Following the cross-validation process, the optimal tuning parameters are assigned to the algorithm to predict cumulative production over time. This

analysis and parameter tuning contributes to the model's accuracy and reliability in forecasting production. This methodology allows us to transfer the selected regression model and its perspective hyperparameters from the scalar model to the vector model.

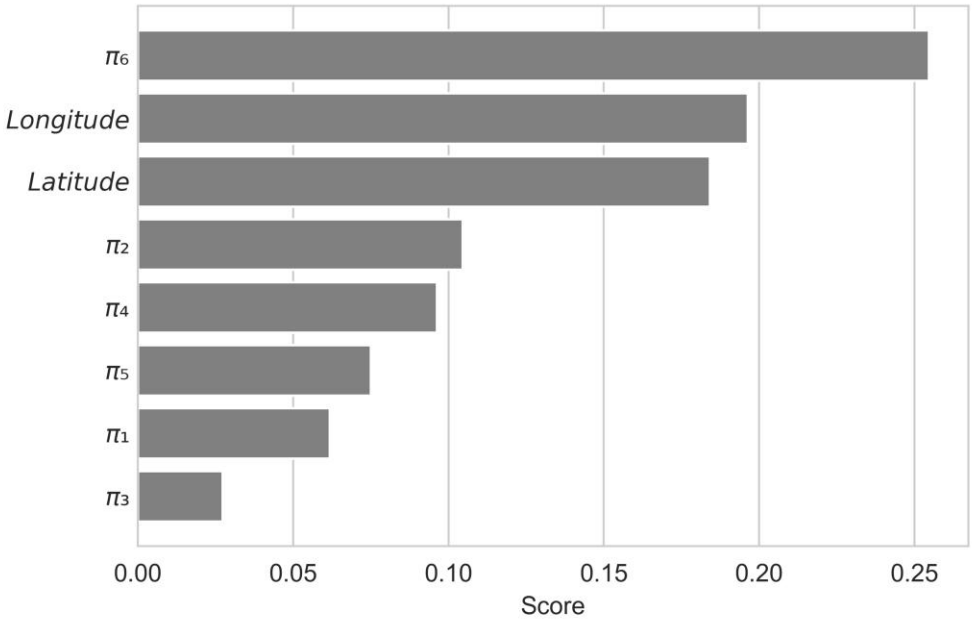

Figure 15. The feature importance ranking for production forecasting shows the top contributing predictive features Π_6, longitude, and latitude, indicating the significant influence of spatial location and selecting dimensionless groups on production.

4.5 Vector Machine Learning Model

Evaluating the ML model's performance in predicting cumulative production curves for undrilled wells in areas of early-stage development poses a challenge due to the absence of actual production data for these specific wells. To assess the model's accuracy, the generated ML cumulative production curves are compared with an oil and gas operator's production curves for undrilled wells, which are vetted by engineers familiar with the

basins from which the data is sourced. This comparison allows for examining the goodness of fit of ML cumulative production curves over a three-year production period.

Incorporating a temporal aspect **Appendix P** into the dimensionless predictors is a crucial step in refining the predictive capabilities of the vector model. By introducing time as a dynamic predictor into the dimensionless groups (e.g. Π_6), ensures that the dimensionless predictors accurately capture the temporal properties of the reservoir behavior over time. This process enhances the model's ability to account for dynamic changes in reservoir conditions and production behavior throughout the production lifespan. This concept allows for the variation in input predictors to influence the vector models predictions over time.

A novel transfer learning method is leveraged to facilitate efficient knowledge transfer and prevent overfitting in predicting cumulative production curves. The regression model and hyperparameters from the scalar (EUR) model, where the 3-year EUR serves as the last point on the cumulative production plot, are transferred to predict the vector (cumulative production curves) model. This transfer learning leverages the knowledge from the trained scalar model to improve efficiency and avoid over-fitting the cumulative production forecasts. Notably, this method demonstrates the versatility of transfer learning, showcasing its applicability beyond traditional deep-learning neural network algorithms to tree-based regression algorithms commonly used in machine learning.

4.5.1 Machine Learning Cumulative Production Forecasting

The training dataset employed to predict the new production curves can significantly influence the hydrocarbon production forecasts, manifesting either negative or positive

impacts. As such, validating the modified dimensionless predictors is important to ensure consistency in the underlying physical processes governing reservoir behavior. This validation process is a crucial step in enhancing the robustness and reliability of the predictive model.

The **appendix Q** provides additional validation of the machine learning model's predictive accuracy by incorporating MAE as an uncertainty measure. While the main analysis demonstrates the model's ability to forecast cumulative production trends, the appendix further reinforces these findings by using MAE-based confidence bounds to quantify prediction reliability. By evaluating how predicted production deviates from actual values, **Figures Q1** and **Q2** illustrate that over 80% of forecasts fall within the $\pm\ MAE$ range, confirming the model's robustness. This approach not only enhances confidence in the ML-generated forecasts but also provides a practical method for operators to assess production uncertainty in real-world applications.

In **Figure 16**, the oil forecasts for a Wolfcamp ML cumulative production curve using the data from both the Montney and Wolfcamp formations for training are observed. The Montney wells historically exhibit lower oil production compared to the Wolfcamp. This discrepancy impacts the model's ability to forecast production precisely when applied across different formations. Especially, when data from lower producing wells is incorporated into the training set, it may bias the model towards underpredicting production in higher producing formations. The impact on the ML production forecasts of the combined data from two different formations is seen as the ML model's cumulative production forecast falls below the oil and gas operator's production curve. The ML model

tends to generalize production trends based on input data. The generalization leads to a conservative production estimate for both formations.

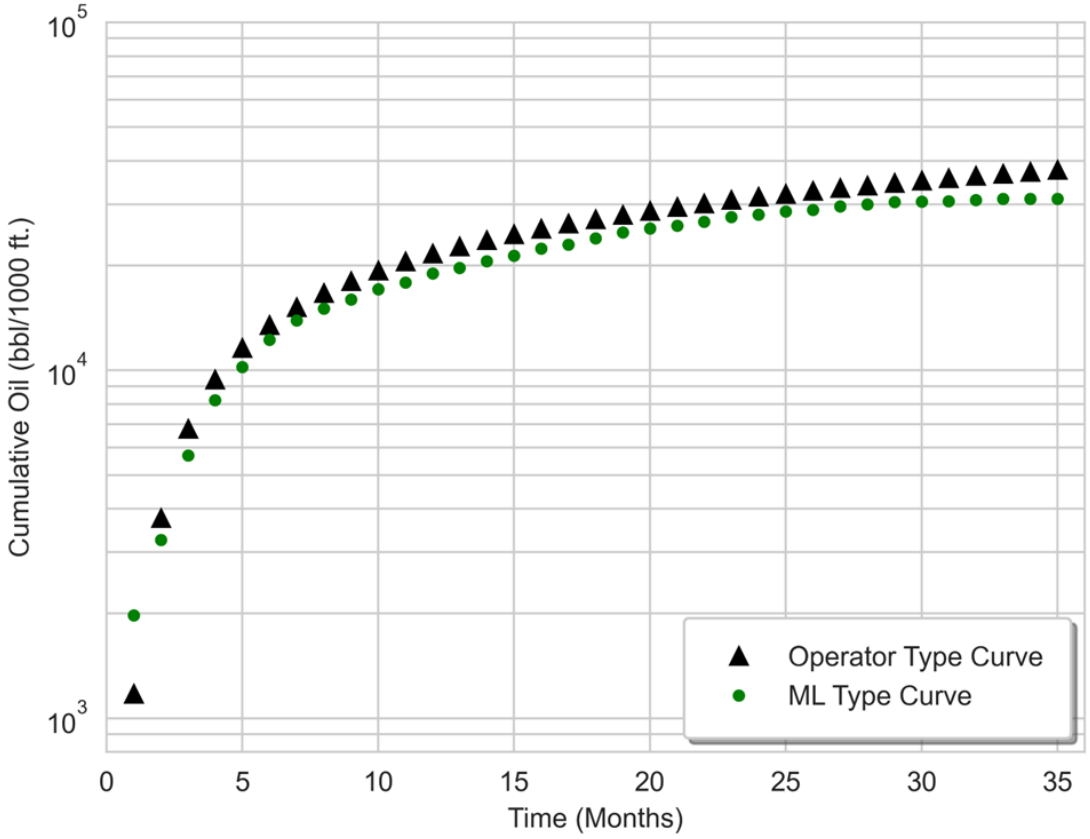

Figure 16. Cumulative oil production forecast for a Wolfcamp well using a ML model trained on both Montney and Wolfcamp data. The ML production curve (green dots) closely follows the operator's production curve (black triangles), but exhibits a slight underprediction, particularly in later months. This suggests that the inclusion of Montney data introduces discrepancies in production behavior, due to geological and completion differences between the formations.

The ML model is then trained using only Wolfcamp formation data. When the Montney production data is removed from the training set, more in tune with the industry practice of using similar wells to forecast undrilled wells, the operator notes that the impact on the

ML production forecasts caused by the Montney wells is alleviated, **Figure 17**. The cumulative production curve provides a very smooth and accurate forecast that aligns well with the oil and gas producer cumulative production curve for undrilled wells in a county near the Wolfcamp training dataset. This further validates the transfer learning of the statistically similar wells in the Wolfcamp outside the new area of interest. Utilizing data from both formations enhances the model's generalization ability, while focusing solely on the formation of interest underscores the efficacy of knowledge transfer.

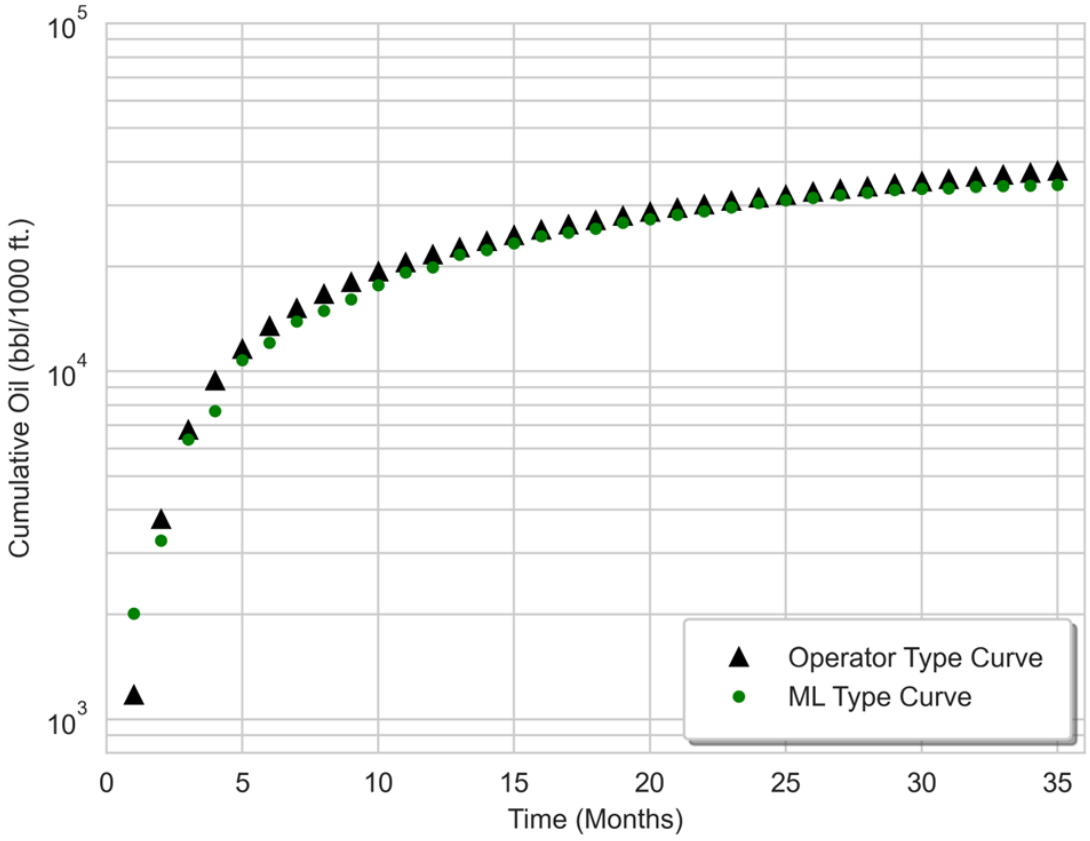

Figure 17. Cumulative oil production forecast for a Wolfcamp well using a machine learning model trained exclusively on Wolfcamp data. The ML production curve (green dots) aligns closely with the operator's production curve (black triangles), demonstrating improved accuracy compared to the mixed-dataset model. This suggests that training on formation-specific data enhances the model's ability to capture unique reservoir characteristics and production behavior.

4.5.2 Validate Transfer Learning – Vector Model

Sensitivity analysis validates knowledge transfer from the scalar to the vector ML model for prediction production trends. The significant role of dynamic input predictor Π_6 in accurate production forecasting, supported by its high feature importance ranking shown in **Figure 15** and corroborated by production forecasts in **Figure 18**, validates its ranked importance and relevance to the model. It underscores $\Pi_6's$ significant influence in shaping predicted production curves. The sensitivity analysis affirms the model's ability to discern and leverage the key variables that contribute to the accuracy of forecasting production trends. This assessment aligns the sensitivity results with the important features.

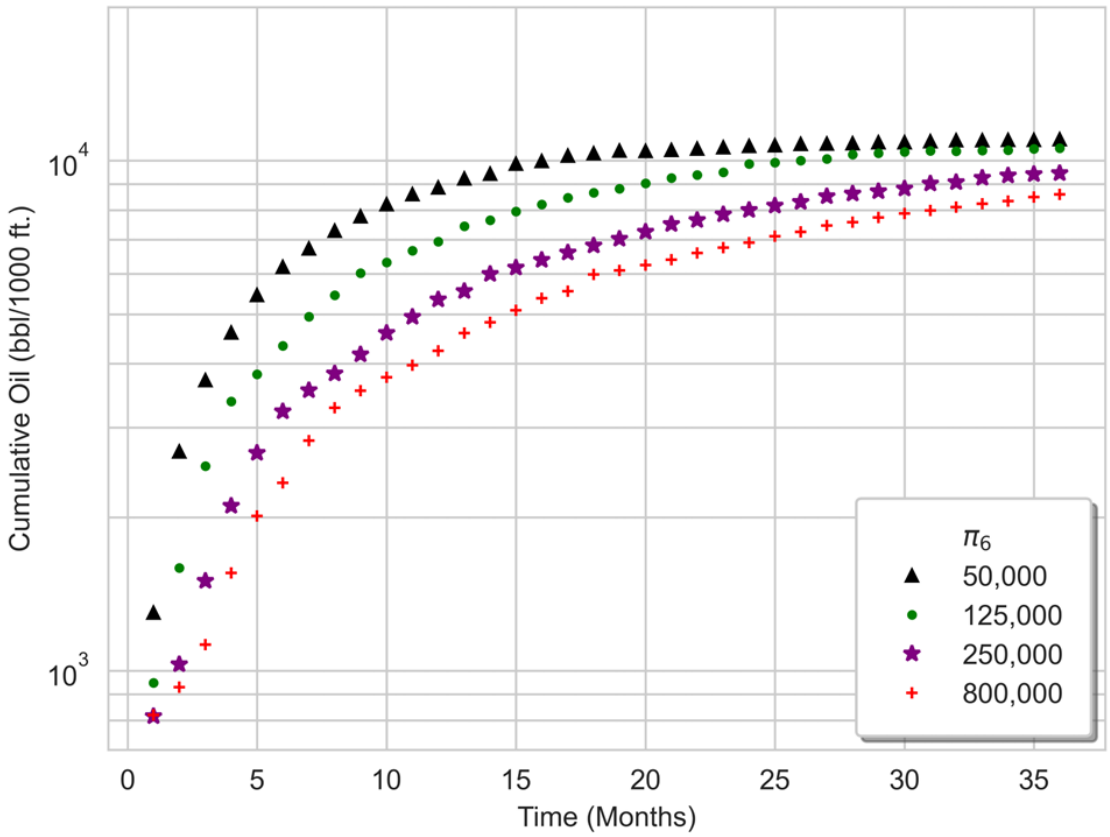

Figure 18. The impact Π_6 on production trends reinforces its significance in improving model forecasts, validating its role as the most influential feature in transfer learning. This result aligns with the feature importance ranking in **Figure 15**, confirming Π_6 as a key variable in shaping production predictions and validating knowledge transfer from scalar to vector ML model.

Conversely, **Figure 19** illustrates the minimal impact of Π_3 on production forecasting. The ML model could be further simplified by removing Π_3 at the operator's discretion. Sensitivity analysis has validated the model's capacity to transfer knowledge from the trained scalar model to the vector model.

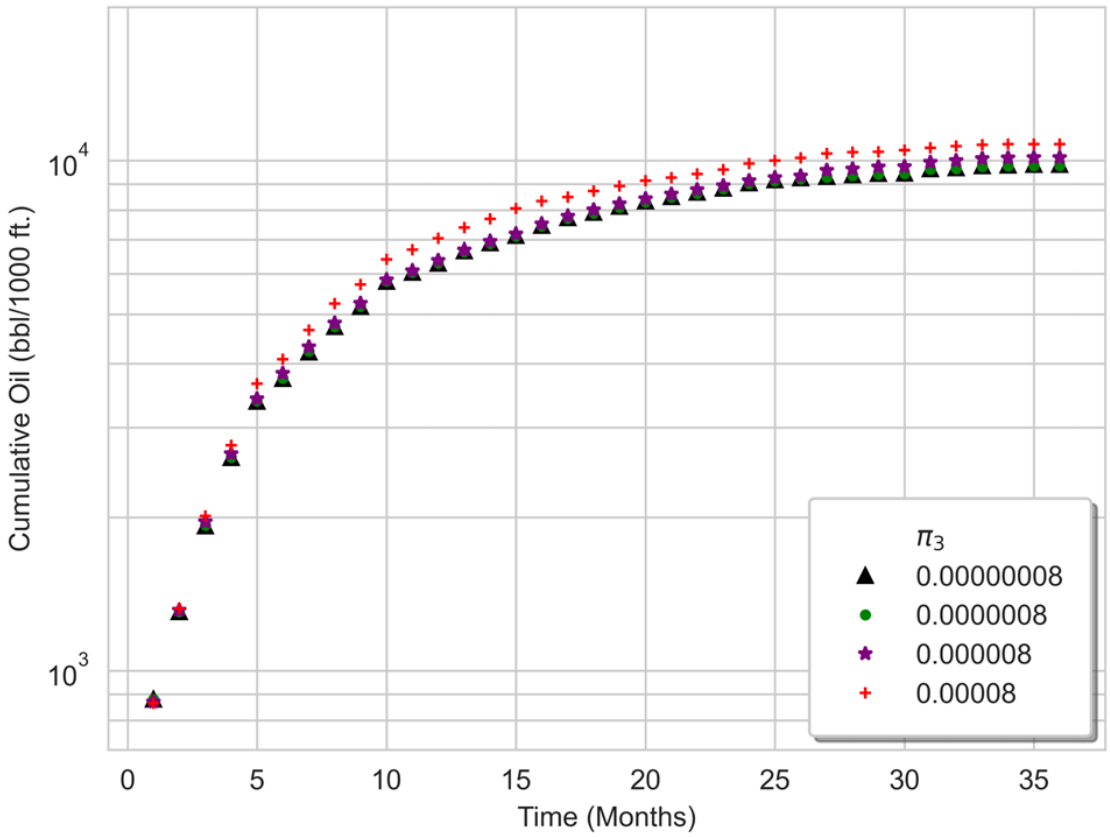

Figure 19. Sensitivity analysis of cumulative oil production per 1000 ft, demonstrating the minimal impact of predictor $\Pi_3$on production forecasting. Variations in Π_3 result in negligible differences in the predicted production curves, suggesting that the model could be further simplified by removing Π_3 without compromising accuracy.

Connecting the Π groups and their respective feature importance results allows for a potential reduction in input predictors, streamlining the model without compromising its accuracy. Variability in feature importance ranking across model iterations was noted due to regression model characteristics.

4.6 Implementing Machine Learning Assisted – Decline Curve Analysis

A novel method, Machine Learning Assisted – Decline Curve Analysis (MLA-DCA), is proposed to address the challenge of generating long-term cumulative production curves

of machine learning. In this method, the machine learning model predicts the first three years of production where training data is available and is extended further into the future using the Arps method. This prediction, combined with the Arps depicted in **Figure 20**, demonstrates the effectiveness in generating accurate production curves with minimal error. The MLA-DCA model highlights the potential of integrating DA, ML, and empirical models to produce long-term forecasts swiftly and accurately. This method provides engineers with a powerful tool capable of producing accurate ML hydrocarbon production forecasts where data is available and empirical forecasts where data is limited or lacking.

The results of this study strongly indicate that the methodologies developed can be readily integrated into the workflow of reservoir evaluation. The ML model exhibits a robust capability to predict EUR. This means that engineers can forecast 3-year EUR, and with an adequate dataset, extend predictions over longer timeframes. The applicability of this work extends to reserve analysis for the development of new fields. Engineers can employ the model to generate production curves for these new reservoirs by leveraging statistical data from similar fields and reservoirs. A key advantage lies in comparing traditionally generated production curves with those produced by the ML model. The model's speed, efficiency, and accuracy provide a valuable tool complementing reservoir and production engineering tasks, enhancing decision making and optimization processes in real world scenarios.

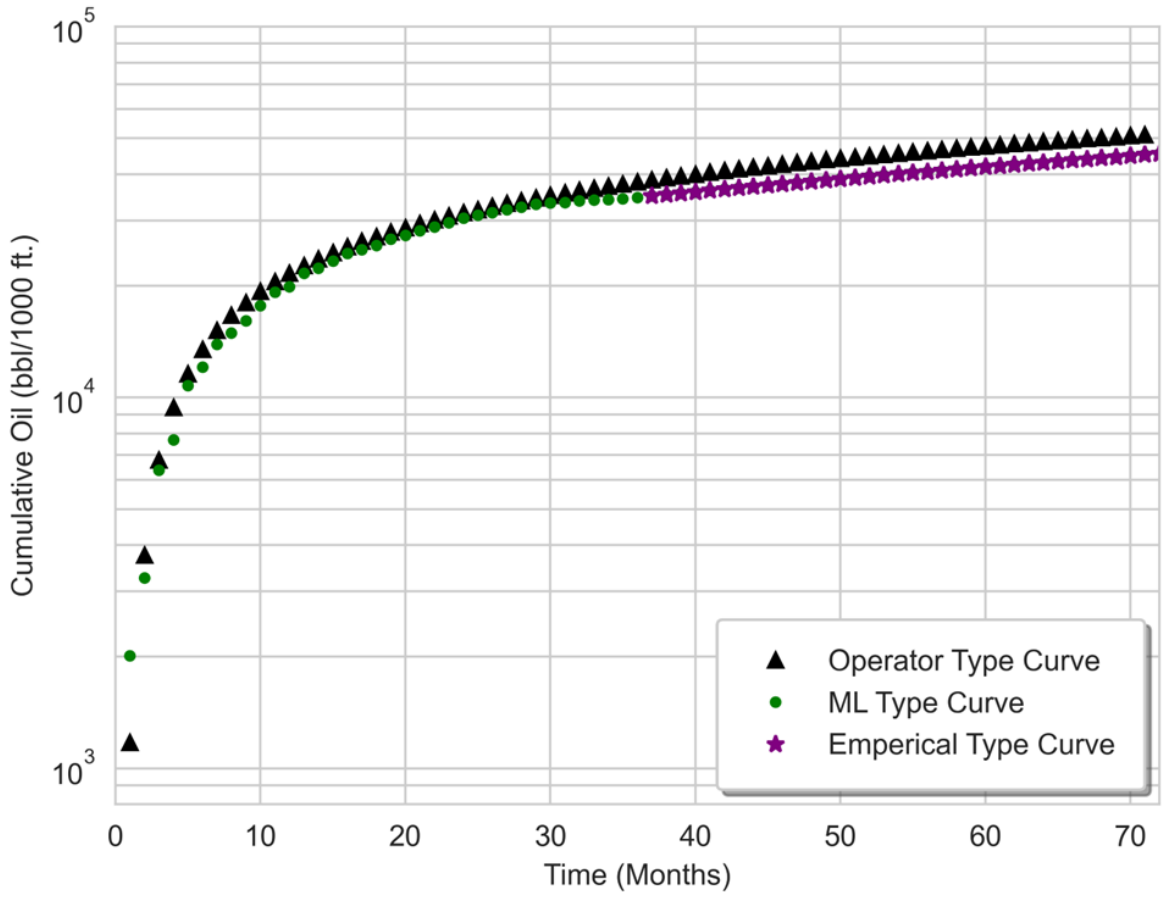

Figure 20. Comparison of cumulative oil production forecasts illustrating the MLA-DCA approach. Here, the ML model (green dots) closely matches the operator's production curve (black triangles) for the initial production phase, while the empirical component (purple stars) extends the forecast into later months. This highlights the power of integrating machine learning and traditional decline curve analysis to generate accurate, long-term production forecasts.

Chapter 5: Conclusions and Future Work

5.1 Conclusions

This study has successfully integrated ML and DA techniques for enhanced hydrocarbon production forecasting. Leveraging transfer learning, the research effectively identifies regression algorithms and hyperparameters, transitioning from scalar to vector predictions. The proposed MLA-DCA method presents an alternate method for predicting long-term production forecasts. The key outcomes and significant findings of this research are summarized as follows:

- Dimensional Analysis Validation: This study demonstrated the suitability and effectiveness of applying DA in conjunction with ML models. DA proves invaluable in reducing dimensionality and simplifying input variables.
- Interpretability through Dimensionless Predictors: Dimensionless predictors derived from DA have reduced the algorithm's complexity and provided meaningful physical insights. This enhances the interpretability of the machine learning model's prediction.
- Transfer Learning: Demonstrated the effectiveness of leveraging a scalar model to determine regression algorithms and hyperparameters for a vector model. Also, transfer learning enables the using statistically similar well data to forecast undrilled wells in areas at the initial stages of development.
- Sensitivity Analysis: This analysis affirms the new model's accurate prediction and responsiveness to input variations, bolstering its reliability and robustness.

- Machine Learning's Predictive Power: The ML models equipped with dimensionless predictors have demonstrated their capability to accurately predict the EUR and cumulative production. This highlights the potential of combining DA and ML in hydrocarbon production forecasting.
- Machine Learning Assisted – Decline Curve Analysis Method for Long-Term Production Forecasting: The MLA-DCA method has provided a practical solution for effectively generating long-term cumulative production curves, analogous to traditional ones. This method combines the strengths of machine learning and decline curve analysis, offering a practical solution for forecasting long-term production trends.

In essence, this study showcases the integration of ML and DA, presenting innovative methodologies to refine the accuracy and reliability of hydrocarbon production forecasts. Transfer learning is a pivotal technique to effectively build an ML model and augment predictions for undrilled wells. Furthermore, the research advocates the MLA-DCA method to extend production forecasting capabilities where training data is lacking. These findings significantly contribute to machine learning and petroleum engineering, offering promising prospects for application across real-world contexts.

5.2 Future Work

This study uses a small dataset to establish and demonstrate the proposed methodology. This approach enabled the expedited acquisition of results and mirrors common industry practices, where production curves are often generated from constrained datasets within specific geographic locations. Nonetheless, this approach inherently limits the

generalizability of the findings to a small subset of wells in analogous fields. Future research endeavors should encompass larger and more diverse datasets to mitigate these constraints and augment the applicability of the methodology. Additionally, while the MLA-DCA method offers improved forecasting accuracy, it remains complex due to the integration of dimensional analysis, machine learning, and empirical models. Simplifying the approach for broader accessibility would enhance its applicability in different settings. Finally, the model's dependence on accurate input data for dimensional analysis and careful transfer learning selection highlights the need for further testing in data-scarce environments to ensure robustness.

Furthermore, to optimize the approach, it is proposed to utilize only completions data without detailed reservoir characterization. Leveraging latitude and longitude coordinates can implicitly account for reservoir characteristics, while depth in the completions data can offer additional insights. Moreover, being more selective in the choice of Π groups could enhance the robustness of the analysis, as evidenced by the diverse range of possible Π groups outlined in **Appendix R**.

Acknowledgement:

Thank you Enverus for your supporting the University of Oklahoma by providing quality data for oil and gas wells.

Nomenclature

D_t	=	Depth from the top	L	ft
g	=	Gravitational acceleration	$\frac{L}{T^2}$	$\frac{ft}{s^2}$
k	=	permeability	L^2	nd
N_c	=	Clusters per 1000 feet		
N_S	=	shots per 1000 ft		
P_i	=	initial reservoir pressure	$\frac{M}{LT^2}$	psi
P_T	=	average treatment pressure	$\frac{M}{LT^2}$	psi
Q_T	=	average treatment rate	$\frac{L^3}{T}$	$\frac{bbl}{min}$
S	=	average stage spacing	L	ft
SG	=	gas specific gravity		
t	=	producing time	T	*T*
V_s	=	slurry volume	L^3	*Bbl*
ρ_o	=	oil density	$\frac{M}{L^3}$	$\frac{lbm}{ft^3}$
ρ_s	=	slurry density	$\frac{M}{L^3}$	$\frac{lbm}{ft^3}$
ϕ	=	Porosity		
μ_T	=	treatment fluid viscosity	$\frac{M}{LT}$	*cp*
μ_o	=	oil viscosity	$\frac{M}{LT}$	*cp*

References

1. Abduvayt, P., Arihara, N., Manabe, R., Ikeda, K. et al. 2003. Experimental and modeling studies for gas-liquid two-phase flow at high pressure conditions, *Journal of the Japan Petroleum Institute*, **46** (2): 111-125.
2. Abduvayt, P., Manabe, R., Arihara, N. et al. 2003. Effects of pressure and pipe diameter on gas-liquid two-phase flow behavior in pipelines. Presented at SPE Annual Technical Conference and Exhibition, 5-8 October. doi:10.2118/84229-MS
3. Agarwal, R., Gardner, D., Kleinsteiber, S. et al. 1999. Analyzing well production data using combined-type-curve and decline-curve analysis concepts. *SPE Reservoir Evaluation and Engineering* **2** (5): 478-486. https://doi.org/10.2118/49222-MS.
4. Agarwal, A., Kenney, A., Tan, Y. et al. 2023. MDI+: A Flexible Random Forest-Based Feature Importance Framework. Berkely, California, USA. arXiv:2307.01932. https://arxiv.org/abs/2307.01932 (preprint: last revised 4 july 2023).
5. Akai, T. and Wood, J.M. 2014. Petrophysical analysis of a tight siltstone reservoir: Montney Formation, Western Canada. Paper presented at the SPWLA 20th Formation Evaluation Symposium of Japan, Chiba, Japan, October 1-2. SPWLA-JFES-2014-E.
6. Al-Safran, E., and Al-Qenae, K. 2018. A study of flow-pattern transitions in high-viscosity oil-and-gas two-phase flow in horizontal pipes. *SPE Production & Operation* **33** (02): doi:10.2118/187939-PA
7. Alghazal, M. A. and Krinis, D. 2023. Data-Driven Modeling of Oil Saturation from Dielectric Logs Using Ensemble Regression, Dimensionality Reduction and Anomaly Detection Machine Learning Algorithms. Paper presented at the ADIPEC, Abu Dhabi, UAE, 2-3 October. https://doi.org/10.2118/216430-ms.
8. Ali, T. and Sheng, J. 2015. Production decline models: A comparison study. Presented at the SPE Eastern Regional Meeting, Morgantown, West Virginia, USA, 13-15 October. https://doi.org/10.2118/177300-MS.
9. Ali, T., Sheng, J., Soliman, M. et al. 2014. New Production-Decline Models for Fractured Tight and Shale Reservoirs. Paper presented at SPE Western North American and Rocky Mountain Joint Meeting, Denver, Colorado, 17-18 April. https://doi.org/10.2118/169537-ms.
10. Amr, S., El-Ashhab, H., El Saban, M. et al. 2018. A Large-Scale Study for a Multi-Basin Machine Learning Model Predicting Horizontal Well Production. Presented at the SPE Annual Technical Conference and Exhibition, Dallas, Texas, USA, 24-26 September. https://doi.org/10.2118/191538-MS.
11. Arps, J.J. 1945. Analysis of decline curves. *Trans.* **160** (01): 228–247. SPE-945228-G. https://doi.org/10.2118/945228-G.
12. Arya, A. and Gould, T.L. Comparison of two phase liquid holdup and pressure drop correlations across flow regime boundaries for horizontal and inclined pipes. *AIME*. Society of Petroleum Engineers.
13. Astarita, G. 1997. Dimensional analysis, scaling, and order of magnitude. *Chemical Engineering Science* **52** (24), 4681-4698. Elsevier. https://doi.org/10.1016/S0009-2509(97)85420-6.
14. Attenborough, K. 2015. *Mathematics and Physics Miscellany*. 2015 Edition. Loughborough University. HELM. Workbook 47. https://nucinkis-lab.cc.ic.ac.uk/HELM/HELM_Workbooks_46-50/WB47-all.pdf.
15. Awad, M. M. 2012. Two-Phase flow. In S. N. Kazi (Ed.), a*n Overview of Heat*

Transfer Phenomena (Chap. 11). Rijeka: InTech. https://doi.org/10.5772/54291
16. Bakay, A., Caers, J., Mukerji, T. et al. 2019. Machine Learning of Spatially Varying Decline Curves for the Duvernay Formation. Presented at the SPE Annual Technical Conference and Exhibition, Calgary, Alberta, Canada, 30 September. https://doi.org/10.2118/196110-MS.
17. Baker, O., Simultaneous flow in oil and gas, *Oil and Gas J*., 53, 185- 195, 1954.
18. Barnea, Dvora, Ovadia Shoham, and Yehuda Taitel. 1982. Flow pattern transition for downward inclined two phase flow; horizontal to vertical." *Chemical Engineering Science* **37** (5): 735-740.
19. Beggs, D. H. and Brill, J. P. 1973. A study of two-phase flow in inclined pipes. *Journal of Petroleum technology*, **25** *(*05): 607-617. doi:10.2118/4007-PA
20. Brito, A., Cabello, R., Marquez, J., Trujillo, J., and Guzman, N. 2015. Flow pattern transition model for gas-highly viscous fluids in horizontal pipelines. In *17th International Conference on Multiphase Production Technology*. BHR Group.
21. R. Balasubramaniam, E. Ramé, J. Kizito, and M. Kassemi 2006. Two Phase Flow Modeling: Summary of Flow Regimes and Pressure Drop Correlations in Reduced and Partial Gravity, NASA/CR
22. Bellman, R. 1961. *Adaptive Control Processes: A Guided Tour*. Princeton University Press. http://www.jstor.org/stable/j.ctt183ph6v.
23. Bhaskar, R. and Nigam, A. 1990. Qualitative Physics Using Dimensional Analysis. *In Artificial Intelligence* **45** (1–2): 73–111. Elsevier BV. https://doi.org/10.1016/0004-3702(90)90038-2.
24. Bridgman, P.W. 1922. *Dimensional Analysis*. First edition. Yale University Press, New Haven. https://archive.org/details/dimensionalanaly00bridrich/page/n9/mode/2up.
25. Bozinovski, S. 2020. Reminder of the First Paper on Transfer Learning in Neural Networks, 1976. *Informatica* **44** (3): https://doi.org/10.31449/inf.v44i3.2828.
26. Breiman, L. 2001. RANDOM FORESTS. *Machine learning* **45:** 5-32**.** California, Berkeley, CA 94720. USA.
27. Buckingham, E. 1914. On physically similar systems; illustrations of the use of dimensional equations. *Physical Review* **4** (4): 345–376. https://doi.org/10.1103/PhysRev.4.345.
28. Cao, Q., Banerjee, R., Gupta, S. et al. 2016. Data Driven Production Forecasting Using Machine Learning. Paper presented at SPE Argentina Exploration and Production of Unconventional Resources Symposium, Buenos Aires, Argentina, 1-3 June. https://doi.org/10.2118/180984-MS.
29. Chang, C. and Lin, C. 2022. LIBSVM: A Library for Support Vector Machines. Department of Computer Science. National Taiwan University. Taipei, Taiwan. *ACM transactions on intelligent systems and technology* **2** (3): 1-27. http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf.
30. Childers, D. and Wu, X. 2020. Forecasting shale gas performance using the Connected Reservoir Storage Model. *Journal of Natural Gas Science and Engineering* **82**: 103499. https://doi.org/10.1016/j.jngse.2020.103499.
31. Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales, *Educational and Psychological Measurement,* Vol. XX, No. 1
32. Cornelio, J., Razak, S., Cho, Y. et al. 2023. Transfer Learning with Prior Data-Driven Models from Multiple Unconventional Fields. *SPE Journal* **28** (05): 2385-

2414. https://doi.org/10.2118/214312-PA.
33. Dalpiaz, David, R for Statistical Learning, 2018
34. Doscher, T.M. and Lechtenberg, H.J. 1980. Scaled Physical Model Studies of the Stream Drive Process. Fossil Energy. DOE. 12075-1.
35. Duong, A. N. 2010. An Unconventional Rate Decline Approach for Tight and Fracture-Dominated Gas Wells. Paper presented at the Canadian Unconventional Resources and International Petroleum Conference, Calgary, Alberta, Canada, 19-21 October. https://doi.org/10.2118/137748-ms.
36. Dupont, E., Vesselinov, V., Burton, E. et al. 2020. Resource Production Forecasting. US Patent 10,859,725 B2.
37. Elman, J.L., 1990. Finding Structure in Time. *Cognitive Science* **14** (2): 179-211. https://doi.org/10.1016/0364-0213(90)90002-E.
38. Falola, Y., Misra, S., Nunez, A. et al. 2023. Rapid High-Fidelity Forecasting for Geological Carbon Storage Using Neural Operator and Transfer Learning. Paper presented at the ADIPEC, Abu Dhabi, UAE, 2-5 October. https://doi.org/10.2118/216135-MS.
39. Fan, C., Chen, M., Wang, X. et al. 2021. A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. *In Frontiers in Energy Research* **9**: 652801. https://doi.org/10.3389/fenrg.2021.652801.
40. Fanchi, J. 2020. Decline Curve Analysis of Shale Oil Production Using a Constrained Monte Carlo Technique. *Journal of Basic Applied and Sciences* **16** (1): 61–67. Set Publishers. https://doi.org/10.29169/1927-5129.2020.16.08.
41. Friedman, J.H. 2002. Stochastic Gradient Boosting. *Computational statistics and data analysis* **38** (4): 367-78. https://doi.org/10.1016/S0167-9473(01)00065-2.
42. Geertsma, J., Croes, G. A., and Schwarz, N. et al. 1956. Theory of Dimensionally Scaled Models of Petroleum Reservoirs. *Transactions of the AIME* **207** (01): 118–12. Society of Petroleum Engineers. https://doi.org/10.2118/539-g.
43. Geurts, P., Ernst, D., Wehenkel, L. et al. 2006. Extremely Randomized Trees. *Machine Learning* **63** (1): 3-42. https://doi.org/10.1007/s10994-006-6226-1.
44. Gokcal, B. 2005. *Effects of High Oil Viscosity on Two-phase Oil-gass Flow Behavior in Horizontal Pipes* Doctoral dissertation. University of Tulsa.
45. Gokcal, B., Al-Sarkhi, A., Sarica, C., and Al-safran, E. M. 2010. Prediction of slug frequency for high-viscosity oils in horizontal pipes. *SPE Projects, Facilities & Construction*, **5** (03): 136-144. https://doi.org/10.2118/124057-PA.
46. Gould, T. L., Tek, M. R., and Katz, D. L. 1974. Two-phase flow through vertical, inclined, or curved pipe. *Journal of Petroleum Technology*, **26** (08): 915-926. doi:10.2118/4487-PA
47. Gu, S., Wang, J., Xue, L. et al. 2022. Deep-Learning-Based Production Decline Curve Analysis in the Gas Reservoir through Sequence Learning Models. *CMES-Computer Modeling and Engineering in Sciences* **131** (3): 1579-1599. https://doi.org/10.32604/cmes.2022.019435.
48. Hastie, T, Tibshirani, R., Friedman, J.H. et al. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Vol. 2. New York NY, USA. Springer.
49. Ilk, D., Rushing, J. A., Perego, A. D., Blasingame, T. A. et al. 2008. Exponential vs. Hyperbolic Decline in Tight Gas Sands — Understanding the Origin and Implications

for Reserve Estimates Using Arps' Decline Curves. Paper presented at the SPE Annual Technical Conference and Exhibition, Denver, Colorado, USA, 21-24 September. https://doi.org/10.2118/116731-ms.
50. James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. an introduction to statistical learning with applications in R., Springer Science+Business Media New York
51. Jayawardena, S. S., Balakotaiah, V., and Witte, L., 1997. Pattern transition maps for microgravity two-phase flow, *AIChE Journal*, **43** (6): 1637-1640.
52. Jha, H. S., Aaditya K., Lee, J. et al. 2022. Statistical and Machine-Learning Methods Automate Multi-Segment Arps Decline Model Workflow to Forecast Production in Unconventional Reservoirs. Paper presented at the SPE Canadian Energy Technology Conference, Calgary, Alberta, Canada, 16 March. https://doi.org/10.2118/208884-MS.
53. Kang, C., Jepson, W. P., and Wang, H. 2002. Flow regime transitions in large diameter inclined multiphase pipelines. In *CORROSION 2002*. NACE International.
54. Karam, M. and Saad, T. 2021. BuckinghamPy:A Python software for dimensional analysis. *SoftwareX* **16**: 100851. https://doi.org/10.1016/j.softx.2021.100851.
55. Karami, A., Mohammadi, A.H., Lee, M. et al. 2017. Decline Curve Based Models for Predicting Natural Gas Well Performance. *Petroleum* **3** (2): 242-248. https://doi.org/10.1016/j.petlm.2016.06.006.
56. 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, *Intl. J. Multiphase Flow* **67**, 37-51.
57. Kokal, S., 1987. An Experimental Study of Two-Phase Flow in Inclined Pipes. Ph.D. Dissertation, University of Calgary, Canada.
58. Kouba, G.E., 1986. Horizontal Slug Flow Modeling and Metering. Ph.D. Dissertation, The University of Tulsa, OK.
59. Kononenko, I. and Kukar, M. 2007. M*achine Learning and Data Mining: Introduction to Principles and Algorithms*. Horwood Publishing Limited, Coll House, Westergate, Chichester, West Sussex, P020 3QL, UK.
60. Kuhn, M. and Johnson, K. 2016. *Applied Predictive Modeling*. Vol. 26. Springer. New York. USA.
61. Labedi, R. 1992. Improved Correlations for Predicting the Viscosity of Light Crudes. *Journal of Petroleum Science and Engineering* **8** (3): 221-234. Elsevier. https://doi.org/10.1016/0920-4105(92)90035-Y.
62. Langhaar, H.L. 1951. *Dimensional Analysis and Theory of Models*. New York, USA: John Wiley and Sons, Inc.
63. Lee, A.L., Gonzalez, M.H., Bertram, E.E. et al. 1966. The Viscosity of Natural Gases. *J Pet Technol* **18** (1966): 997-1000. https://doi.org/10.2118/1340-PA.
64. Lee, J. and Wattenbarger, R.A. 1996. Gas Reservoir Engineering. *SPE Textbook Series* **5**. Society of Petroleum Engineers. USA.
65. Leverett, M.C., Lewis, W.B., True, M.E. et al. 1942. Dimensional-model Studies of Oil-field Behavior. *Transactions of the AIME* **146**. https://doi.org/10.2118/942175-G.
American Conference on Multiphase Production Technology. BHR Group.
66. Li, B., Billiter, T., Tokar, T. et al. 2021. Rescaling Method for Improved Machine-Learning Decline Curve Analysis for Unconventional Reservoirs. *SPE Journal* **26** (04): 1759-1772. https://doi.org/10.2118/205349-PA.
67. Li, W., Dong, Z., Lee, J. et al. 2022 Development of Decline Curve Analysis

Parameters for Tight Oil Wells Using a Machine Learning Algorithm. *Geofluids* **2022**. 8441075. https://doi.org/10.1155/2022/8441075.
68. Li, Y. and Han, Y. 2017. Decline Curve Analysis for Production Forecasting Based on Machine Learning. Presented at the SPE Symposium: Production Enhancement and Cost Optimisation, Kuala Lumpur, Malaysia, 7-8 November. https://doi.org/10.2118/189205-MS.
69. Li, X., Miskimins, J. L., Sutton, R. P., and Hoffman, B. T. 2014. Multiphase flow pattern recognition in horizontal and upward gas-liquid flow using support vector machine models. In *SPE Annual Technical Conference and Exhibition*. doi:10.2118/170671-MS
70. Lin, P., 1982. Flow Regime Transitions In Horizontal Gas–Liquid Flow. M.S. Thesis, University of Illinois, IL.
71. Liu, H., Zhang, J., Liang, F. et al. 2021. Incorporation of Physics into Machine Learning for Production Prediction from Unconventional Reservoirs: A Brief Review of the Gray-Box Approach. *SPE Res Eval and Eng* **24** (2021): 847-854. https://doi.org/10.2118/205520-PA.
72. Loomis, A. G. and Crowell, D. C. 1964. Theory and Application of Dimensional and Inspectional Analysis To Model Study of Fluid Displacements in Petroleum Reservoirs. U.S. Bureau of Mines. RI6546.
73. Manabe, R., 2001. A Comprehensive Mechanistic Heat Transfer Model For Two-Phase Flow With High-Pressure Flow Pattern Validation. Ph.D. Dissertation, The University of Tulsa, OK.
74. Mandhane, J. M., Gregory, G. A. and Aziz, K. 1974. A Flow Pattern Map for Gas-Liquid Flow in Horizontal Pipes: Predictive Models, *Int. J. Multiphase Flow*, 1, 537-553.
75. Mask, G., Wu, X., Ling, K. et al. 2019. An improved model for gas-liquid flow pattern prediction based on machine learning. *Journal of Petroleum Science and Engineering.* **183**: 106370. https://doi.org/10.1016/j.petrol.2019.106370.
76. Mask, G. and Wu, X. 2023. Deriving New Type Curves through Machine Learning in the Wolfcamp Formation. Presented at the SPE Reservoir Characterisation and Simulation Conference, 24-26 January. SPE-212624-MS. https://doi.org/10.2118/212624-MS.
77. Mask, G. and Wu, X. 2023. Generating New Type Curves through Machine Learning Utilizing Dimensional Analysis. Presented at the SPE Oklahoma City Oil and Gas Symposium. 19-22 April. https://doi.org/10.2118/213080-MS.
78. Mask, G., Wu, X., Nicholson et al. 2025. Enhanced hydrocarbon production forecasting combining machine learning, transfer learning, and decline curve analysis. *Gas Science and Engineering* **134**: https://doi.org/10.1016/j.jgsce.2024.205522
79. Mata, C., Pereyra, E., Trallero, J.L., Joseph, D.D., 2002. Stability of stratified gas–liquid flows. International Journal of Multiphase Flow 28, 1249–1268.
80. McCulloch, W.S. and Pitts, W.H. 1943. A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*. **5**: 115-133. https://doi.org/10.1007/BF02478259.
81. Meng, W., 2001. Low Liquid Loading Gas–Liquid Two-Phase Flow in Near-Horizontal Pipes. Ph.D. Dissertation, The University of Tulsa, OK.
82. Mireault, R. and Dean, L. 2008. Reservoir engineering for geologists. Canadian society of petroleum geologists magazine, Calgary, Albert.
83. Mishra, S. 2014. Exploring the Diagnostic Capability of RTA Type Curves. Paper

presented at the SPE Annual Technical Conference and Exhibition, Amsterdam, The Netherlands, 27 October. https://doi.org/10.2118/173481-STU.
84. Misra, S., Elkady, M., Kumar, V. et al. 2024. Use of Transfer Learning in Shale Production Forecasting. Paper presented at the International Petroleum Technology Conference, Dhahran, Saudi Arabia, 14 February. https://doi.org/10.2523/IPTC-23438-MS.
85. Mitchell, T.M. 1997. *Machine Learning*. New York, USA: McGraw-Hill Science.
86. Mohaghegh, S. D. and Abdulla, F. 2014. Production Management Decision Analysis Using AI-Based Proxy Modeling of Reservoir Simulations – A Look-Back Case Study. SPE Annual Technical Conference and Exhibition, Amsterdam, The Netherlands, 27-29 October. https://doi.org/10.2118/170664-ms.
87. Morales-German, G., Navarro-Rosales, R., Dubost, F. X. et al. 2012. Production Forecasting for Shale Gas Exploration Prospects Based on Statistical Analysis and Reservoir Simulation. Paper presented at the SPE Latin America and Caribbean Petroleum Engineering Conference, Mexico City, Mexico, 16-18 April. https://doi.org/10.2118/152639-ms.
88. Moradzadeh, A., Mansour Saatloo, A., Mohammadi-ivatloo, B. et al., 2020. Performance Evaluation of Two Machine Learning Techniques in Heating and Cooling Loads Forecasting of Residential Buildings. *Applied Sciences*. **10**. https://doi.org/10.3390/app10113829.
89. Nicholson, C. 2018. Data Preparation. Lecture. University of Oklahoma. Norman, OK, USA.
90. Nieto, J., Bercha, R., Chan, J. et al. 2009. Shale Gas Petrophysics - Montney and Muskwa, Are They Barnett Look-Alikes?. SPWLA-2009-84918. SPWLA Annual Logging Symposium, The Woodlands, Texas, 21-24 June.
91. Oddie, G., Shi, H., Durlofsky, L.J., Aziz, K., Pfeffer, B., and Holmes, J.A. 2003. Experimental study of two and three phase flows in large diameter inclined pipes. *International Journal of Multiphase Flow*, 29, 527-558.
92. Odi, U., Ayeni, K., Alsulaiman, N. et al. 2021. Applied Transfer Learning for Production Forecasting in Shale Reservoirs. Paper presented at the SPE Middle East Oil and Gas Show and Conference, event canceled, 28 November. https://doi.org/10.2118/204784-MS.
93. Ojeda, L. 2023. Application of Machine Learning and Dimensional Analysis to Evaluate the Performances of Various Downhole Centrifugal Separator Types. Paper presented at the SPE Annual Technical Conference and Exhibition, San Antonio, Texas, USA, 16-18 October. https://doi.org/10.2118/213059-MS.
94. Ojha, S. P., Misra, S., Tinni, A., Sondergeld, C., Rai, C. et al. 2017. Relative permeability estimates for Wolfcamp and Eagle Ford shale samples from oil, gas and condensate windows using adsorption-desorption measurements. *Fuel* **208**: 52–64. Elsevier BV. https://doi.org/10.1016/j.fuel.2017.07.003.
95. Omebere-Iyari, N. K., Azzopardi, B. J. and Ladam, Y. 2007. Two-phase flow patterns in large diameter vertical pipes at high pressures. *AIChE J*., 53, 2493-2504. doi:10.1002/aic.11288
96. Pan, S.J. and Yang, Q. 2010. A Survey on Transfer Learning. *IEEE Transaction on Knowledge and Data Engineering*, **22** (10), 1345-1359. https://doi.org/10.1109/TKDE.2009.191.

97. Pearson, K. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2** (11): 559–572. Informa UK Limited. https://doi.org/10.1080/14786440109462720.
98. Pedregosa, F., Varoquaux, G., Gramfort, A. et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12** (12): 2825-2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.
99. Pereyra, E., Torres, C., Mohan, R., Gomez, L., Kouba, G., and Shoham, O. (2012). A methodology and database to quantify the confidence level of methods for gas–liquid two-phase flow pattern prediction. *Chemical Engineering Research and Design*, *90*(4), 507-513.
100. Peter, S.C, Dhanjal, J.K., Malik, V. et al. 2019. Quantitative Structure-Activity Relationship (QSAR): Modeling Approaches to Biological Applications. *Encyclopida of Bioinformatics and Computational Biology*. **2**: 661-676, https://doi.org/10.1016/B978-0-12-809633-8.20197-0.
101. Platt, J.C. 2000. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers*. **10**.
102. Peters, E.J. 2001. *Advanced Petrophysics* Texas, USA. Greenleaf Book Group.
103. Pyrak-Nolte, L. J., Giordao, N., Nolte, D. D. et al. 2004. Experimental Investigation of Relative Permeability Upscaling from the Micro-Scale to the Macro-Scale. Office of Scientific Information. https://doi.org/10.2172/833410.
104. Rahmanifard, H., Gates, I., Shabib-Asl., A. et al. 2022. Comparison of Machine Learning and Statistical Predictive Models for Production Time Series Forecasting in Tight Oil Reservoirs. Paper presented at the SPE/AAPG/SEG Unconventional Resources Technology Conference, Houston, Texas, USA, 20-22 June. https://doi.org./10.15530/urtec-2022-3703284.
105. Rosenblatt, F. 1958. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*. **65** (6): 386-408. https://doi.org/10.1037/H0042519.
106. Rosenblatt, F. 1961. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. https://doi.org/10.2307/1419730.
107. Ruark, A.E. 1935. Inspectional Analysis: A Method Which Supplements Dimensional Analysis. *Journal of the Elisha Mitchell Scientific Society* **51** (1): 127-133. North Carolina Academy of Sciences, Inc. https://www.jstor.org/stable/24332065.
108. Saleh, H. 2022. *Machine Learning – Regression*. MS Thesis, Higher Institute for Applied Sciences and Technology, Syrian Arab Republic (December 2022). https://doi.org/10.13140/RG.2.2.35768.67842.
109. Sanders, M., Felling, K., Thomson, S., Wright, S., Thorpe, R. et al. 2016. Dry Polyacrylamide Friction Reducer: Not just for Slick Water. Paper presented at the SPE Hydraulic Fracturing Technology Conference, The Woodlands, Texas, USA, 9-11 February. https://doi.org/10.2118/179146-MS.
110. Sarduy, L. A. and Edelmann, U. C. 2018. Method for Estimating Oil/Gas Production Using Statistical Learning Models. European Patent No. EP3414428A2.
111. Sarker, I.H. 2021. Machine Learning: Algorithms, Real-World Applications and Research Direction. *SN Computer Science* **2** (3): 2661-8907.

https://doi.org/10.1007/s42979-021-00592-x.
112. Shoham, O., 2006. Mechanistic Modeling of Gas–Liquid Two-Phase Flow in Pipes. Society of Petroleum Engineers, ISBN 978-1-55563-107-9.
113. Shoham, O., 1982. Flow Pattern Transitions and Characterization in Gas–Liquid Two Phase Flow in Inclined Pipes. Ph.D. Dissertation, Tel-Aviv, University, Israel.
114. Sonin, A. 2001. *The Physical Basis of Dimensional Analysis.* Second edition. Cambridge, MA.
115. Stephenson, S.V. 1999. Method of Controlling Development of an Oil and Gas Reservoir. US Patent 6002985A.
116. Taitel, Y. and Dukler, A.E. 1976. A model for predicting flow regime transitions in horizontal and near horizontal gas-liquid flow. *AIChE Journal*, 22(1), 47-55.
117. Tan, C., Sun, F., Kong, T. et al. 2018. A Survey on Deep Transfer Learning. *Artificial Neural Networks and Machine Learning*. **11141**. https://doi.org/10.1007/978-3-030-01424-7_27.
118. Thompson, E. and Ripperger, E.A. 1964. An Experimental Technique For The Investigation Of The Flow Of Halite And Sylvinite. U.S. Rock Mechanics/Geomechanics Symposium. ARMA-64-467.
119. Vaisblat, N., Harris, N. B., Ayranci, K., Power, M., DeBhur, C., Bish, D. L., Chalaturnyk, R., Krause, F., Crombez, V., Euzen, T., Rohais, S. et al. 2021. Compositional and diagenetic evolution of a siltstone, with implications for reservoir quality; an example from the Lower Triassic Montney Formation in western Canada. *Marine and Petroleum Geology* **129** (2021). Elsevier BV. https://doi.org/10.1016/j.marpetgeo.2021.105066.
120. Vaisblat, N., Harris, N. B., Ayranci, K., Chalaturnyk, R., Power, M., Twemlow, C., and Minion, N. et al. 2022. Petrophysical properties of a siltstone reservoir - An example from the Montney Formation, western Canada. *Marine and Petroleum Geology* **136**: 105431. Elsevier BV. https://doi.org/10.1016/j.marpetgeo.2021.105431.
121. Valkó, P.P. and Lee, W.J. 2010. A better way to forecast production from unconventional gas wells. Paper presented at the SPE Annual Technical Conference and Exhibition, Florence, Italy, 19-22 September. https://doi.org/10.2118/134231-MS.
122. Valkó, P. P. 2009. Assigning value to stimulation in the Barnett Shale: a simultaneous analysis of 7000 plus production Histories and Well completion records. Paper presented at the SPE Hydraulic Fracturing Technology Conference, The Woodlands, Texas, 19 January. https://doi.org/10.2118/119369-ms.
123. Van Dresar, N., Siegwarth, J., 2001. Near-Horizontal, Two-Phase Flow Patterns of Nitrogen and Hydrogen at Low Mass and Heat Flux. National Aeronautics and Space Administration (NASA), http://gltrs.grc.nasa.gov/GLTRS. Vapnik, V. N., 2000. *The Nature of Statistical Learning Theory.* New York, USA. Springer.
124. Vapnik, V. N., 2000. *The Nature of Statistical Learning Theory.* New York, USA. Springer.
125. Vega-Ortiz, C., Panja, P., Deo, M. et al. 2023. Decline Curve Analysis Using Machine Learning Algorithms: Rnn, Lstm, and Gru. Paper presented at the 57th U.S. Rock Mechanics/Geomechanics Symposium, Atalanta, Georgia, USA, 25-28 June. https://doi.org/10.56952/ARMA-2023-0287.
126. Vieira, C., Kallager, M., Vassmyr, M., La Forgia, N., and Yang, Z. 2018. Experimental investigation of two-phase flow regime in an inclined pipe. In *11th North*

127. Vocke, C. P., Deglint, H. J., Clarkson, C. R. et al. 2018. Estimation of Petrophysical Properties of Tight Rocks from Drill Cuttings Using Image Analysis: An Integrated Laboratory-Based Approach. Paper presented at the SPE Canada Unconventional Resources Conference, Calgary, Alberta, Canada, 13-14 March. https://doi.org/10.2118/189825-ms.
128. Vyas, A., Datta-Gupta, A., Mishra, S. et al. 2017. Modeling Early Time Rate Decline in Unconventional Reservoirs Using Machine Learning Techniques. Paper presented at the Abu Dhabi International Petroleum Exhibition and Conference, Abu Dhabi, UAE, 13-16 November. https://doi.org/10.2118/188231-MS.
129. Wang, M., Huang, D., Wang, G. et al. 2020. SS-XGBoost: A Machine Learning Framework for Predicting Newmark Sliding Displacements of Slopes. *Journal of Geotechnical and Geoenvironmental Engineering* **146** (9): 04020074. https://doi.org/10.1061/(ASCE)GT.1943-5606.0002297.
130. Wang, Y. and Sun, S. 2016. Direct Calculation of Permeability by High-Accurate Finite Difference and Numerical Integration Methods. *Communications in Computational Physics*. **20** (2): 405-440. https://doi.org/10.4208/cicp.210815.240316a.
131. Weiss, K., Khoshgoftaar, T.M., Wang, D. et al. 2016. A survey of transfer learning. *J Big Data*. **3** (9). https://doi.org/10.1186/s40537-016-0043-6.
132. Williams-Kovacs, J. D. and Clarkson, C.R. 2011. Using Stochastic Simulation To Quantify Risk and Uncertainty in Shale Gas Prospecting and Development. Paper presented at the Canadian Unconventional Resources Conference, Calgary, Alberta, Canada, 15-17 November. https://doi.org/10.2118/148867-MS.
133. White, F.M. 2019. *Fluid Mechanics*. Eighth ed. McGraw-Hill Education. 2 Penn Plaza, New York, NY, USA.
134. Wilkens R., 1997. Prediction of the Flow Regime Transitions In High Pressure Large Diameter, Inclined Multiphase Pipelines. Ph.D. Dissertation, Ohio University, OH.
135. Willmott, C. and Matsuura, K. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *In Climate Research* **30** (1): 79–82. https://doi.org/10.3354/cr030079.
136. Woo, J., Lee, H. S., Ozyer, C. et al. 2021. Effect of Lamination on Shale Reservoir Properties: Case Study of the Montney Formation, Canada. *Geofluids* **2021**: 1-14. https://doi.org/10.1155/2021/8853639.
137. Xue, W., Wang, Y., Wang, T. et al. 2024. Efficient hydraulic and thermal simulation model of the multi-phase natural gas production system with variable speed compressors. *Applied Thermal Engineering* **242**: https://doi.org/10.1016/j.applthermaleng.2024.122411.
138. Yang, L., Hanneke, S., Carbonell, J. et al. 2013. A Theory of Transfer Learning with Applications to Active Learning. *Machine learning* **90** (2013): 161-89. https://doi.org/10.1007/s10994-012-5310-y.
139. Yehia, T., Khattab, H., Tantawy, M., and Mahgoub, I. et al. 2022. Removing the Outlier from the Production Data for the Decline Curve Analysis of Shale Gas Reservoirs: A Comparative Study Using Machine Learning. *ACS Omega* **7** (36): 32046–32061. American Chemical Society. https://doi.org/10.1021/acsomega.2c03238.

Appendix A. The principle of absolute significance of relative magnitude

Understood mathematically, measure the lengths of two objects l_1 and l_2 in feet, then measure the same two objects l_1' and l_2' in meters. The principle of absolute significance of relative magnitude dictates that the ratio is constant:

$$\frac{l_1}{l_2} = \frac{l_1'}{l_2'} \qquad \text{(A-1)}$$

Bridgman obtained a monomial formula comprised of numerical values of base quantities and fulfills the principle of absolute significance of relative magnitude only in the power-law form (Bhaskar and Nigam 1990):

Bridgman's Product Theorem: Let a secondary quantity be derived from measurements of primary quantities α, β, γ, …; assuming absolute significance of relative magnitudes, the value of the secondary quantity is derived as:

$$f = C_1 \alpha^a \beta^b \gamma^c \ldots, \qquad \text{(A-2)}$$

where $C_1, a, b, c\ldots$ are mathematical constants and $\alpha, \beta, \gamma\ldots$ are base quantities (Bridgman 1922).

Assuming the primary quantities α, β, γ, … are fundamental, primary dimensions M, L, T can be related with them (Bhaskar and Nigam, 1990). Using Bridgman's Product Theorem, the secondary quantities of the derived dimensions will be conveyed in the power-law form:

$$f = C_1 M^a L^b T^c \ldots, \qquad \text{(A-3)}$$

Appendix B. Buckingham Π Theorem

Buckingham (1914) states that based on the principle of dimensional homogeneity, a relation subsists among any number of physical quantities of n (independent variable) of different kinds. If several independent variables are of one kind, then let them be represented as r (ratio of any one like quantity and the others of like kinds to this one) (Buckingham, 1914). Let $Q_1, Q_2, \ldots Q_n$ represent one quantity of each kind and the remaining quantities of each kind are specified by their ratios r', r'', ..., etc., to the particular quantity of that kind selected equation is (Buckingham, 1914):

$$f(Q_1, Q_2, \cdots Q_n, r', r'' \cdots) = 0 \qquad \text{(B-1)}$$

If the ratios do not vary the complete set of independent variables is represented as:

$$f(Q_1, Q_2, \cdots Q_n) = 0 \qquad \text{(B-2)}$$

any equation which describes this relationship is reducible to the form:

$$\psi(\Pi_1, \Pi_2, \cdots \Pi_n, r', r'' \cdots) = 0 \qquad \text{(B-3)}$$

If k is the number of fundamental units (e.g., M, L, T) required in an absolute system for measuring the n kinds of quantity, the number of dimensionless products Π is (Buckingham 1914):

$$i = n - k \qquad \text{(B-4)}$$

Therefore:

$$\Pi_1 = \varphi(\Pi_2, \Pi_3, \cdots \Pi_i, r', r'', r''' \cdots) \qquad \text{(B-5)}$$

$$or\ into\ the\ form$$

$$r' = \varphi(\Pi_1, \Pi_2, \Pi_3, \cdots\ \Pi_i, r'', r''' \cdots) \qquad \text{(B-6)}$$

If the ratios do not vary the simplified equation is:

$$\Pi_1 = \varphi(\Pi_2, \Pi_3, \cdots\ \Pi_i) \qquad \text{(B-7)}$$

Buckingham Π Theorem yields a complete set of independent dimensionless Π groups that are not unique.

Appendix C. Dimensionless Decline Curve Analysis Equations

To facilitate analysis, the dimensionless form of the PLE model is given as:

$$q_D(t_D) = exp(-D_\infty t_D - t_D^n) \qquad \text{(C-1)}$$

where:

$q_D = -\frac{q}{q_i}$ is the dimensionless production rate,

$t_D = D_i t$ is the dimensionless time.

This expression combines the exponential decay governed by the asymptotic decline rate D_∞ with a power-law term that captures early-time transient behavior. The use of dimensionless parameters enables the PLE model to generalize across varying reservoir conditions by focusing on relative production behavior rather than absolute values.

To facilitate analysis, the dimensionless form of the SEPD model is given as:

$$q_D(t_D) = exp(-t_D^n) \qquad \text{(C-2)}$$

where:

$q_D = -\frac{q}{q_i}$ is the dimensionless production rate,

$t_D = \frac{t}{\tau}$ is the dimensionless time.

Here, the production decline is expressed as a stretched exponential function, where the exponent n governs the rate of decay and τ serves as a characteristic time scale. This formulation enables the SEPD model to flexibly describe a wide range of decline behaviors by adjusting only a few parameters.

To facilitate analysis, the dimensionless form of Duong's model is given as (Duong 2010; Ali and Sheng 2015):

$$\frac{q_D}{G_D} = at_D^{-m} \qquad \text{(C-3)}$$

This expression relates the dimensionless production rate to the dimensionless cumulative production, highlighting the transient behavior dominated by fracture flow during early time.

$$q_D(t_D) = t_D^{-m} exp\left(\frac{a}{1-m}\left(t_D^{(1-m)} - 1\right)\right) \qquad \text{(C-4)}$$

This is the full time-domain solution for Duong's model, capturing both transient and late-time behavior. It incorporates a combination of power-law and exponential terms to reflect evolving fracture connectivity and depletion effects.

$$t_D(a,m) = t_D^{-m} exp\left(\frac{a}{1-m}(t_D^{1-m} - 1)\right) \qquad \text{(C-5)}$$

This alternative form expresses production rate evolution in terms of model-specific parameters a and m, facilitating sensitivity analysis.

$$q_D(t_D) = q_D^1 t_D(a,m) + q_D^\infty \qquad \text{(C-6)}$$

This expression generalizes Duong's model by introducing scaling and offset terms, q_D^1 and q_D^∞, to better fit varying production decline behaviors across different reservoirs.

where:

q = the production at a given time,

q_i = the initial production rate,

G_p = cumulative production,

G_r = the reference cumulative production,

t_r = the reference time or characteristic time,

q_D^1 and q_D^∞ are scaling constants for production rate behavior.

This normalization framework enables direct comparison of decline models and facilitates their application across wells with different operational scales, reservoir properties, and completion designs.

Appendix D. Empirical Hydrocarbon Production Comparison

Table D1. The comparative analysis of empirical DCA models, highlighting their strengths and weaknesses. Understanding these trade-offs is essential for selecting the most suitable model for unconventional reservoirs (Yehia et al., 2022).

Model	Pros	Cons
Arps (1945)	Simple to apply	Overestimate EUR; Neglects flow behavior of shale gas
MLM (1956)	Applied to depletion reservoirs with low pressure where gravity is the only driving mechanism	Neglects flow behavior of shale gas
K (1970)	Simple to apply	Underestimate EUR; Neglects flow behavior of shale gas
Weng (1984)	Simple with two fitting parameters	Underestimate EUR; Neglects flow behavior
Weibull (1995)	Simple with two fitting parameters	Underestimate EUR; Neglects flow behavior
boundary "b" approach (2000)	Considers flow behavior of shale gas	B value depends on the fluid properties and pressure
Hsieh (2001)	Considers flow behavior of shale gas and reservoir properties	Underestimate EUR
Multistage approach (2007)	Consider flow behavior of shale; Arps structure	Switch point difficult to determine (D_{switch} to t_{switch})
Modified Arps (2008)	Consider flow behavior of shale gas	Multiple steps
PLE (2008)	Developed to match shale gas behavior	Underestimate EUR; 4 fitting parameters being regressed on
T-model (2009)	Applicable to shale gas; Consider flow behavior	Overestimate EUR
SEPD (2010)	Applicable to shale gas; Consider flow behavior	Overestimate EUR; Requires large dataset
Duong (2010, 2011)	Applicable to shale gas; Consider flow behavior	Model parameters overestimated in water breakthrough case
LGM (2011)	Consider flow behavior of shale gas; Simple to apply	Overestimate EUR; Requires large dataset
EEDCA (2015)	Applicable to early and late time production	Overestimate EUR; β_1 fixed value not effecting curve
FDC (2016)	Applicable to shale gas production behavior particularly to late time	Overestimate EU
Wang (2017)	Based on Duong's assumptions; more general	Overestimate EUR
VDMA (2018)	Applicable to shale gas production behavior; Simple	Overestimate EUR

Appendix E. Machine Learning Expressions

To formalize the machine learning models used in this study, the following section presents the mathematical formulations that define each algorithm's structure and behavior. These equations support the implementation and interpretation of model predictions for both flow pattern classification and hydrocarbon production forecasting.

The Linear Regression model is one of the foundational algorithms in supervised learning. It assumes a linear relationship between the input features and the output target variable:

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \qquad \text{(E-1)}$$

where:

$f(x)$ is the derived function used to predict the target variable,

$X \in \mathbb{R}^n$ denotes the $n > 0$ input variables,

β_0 is the intercept,

$\beta_1, \ldots, \beta_n$ denote the regression coefficients,

ϵ is the residual error term.

Linear regression serves as a baseline for model comparison in this study due to its simplicity and interpretability.

In ensemble methods like Random Forest and Extremely Randomized Trees, multiple decision trees are built and their outputs averaged for regression tasks. The general prediction for an ensemble of T trees is: (Breiman, 2001; Pedregosa et al., 2011):

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \qquad \text{(E-2)}$$

where:

$\hat{y}$ is the predicted value,

T is the number of trees in the ensemble,

$h_t(x)$ is the prediction of the t-th tree for the input x.

$h_t(x)$ is the prediction of the t-th tree for the input x.

Gradient Boosting builds trees sequentially, where each tree learns from the residuals of the previous ensemble to iteratively refine the prediction. At each step m, the model updates the predicted value by adding the output of a newly fitted tree to the previous model (Friedman, 2002) as follows:

$$\hat{y}_m(x) = \hat{y}_{m-1}(x) + \gamma h_m(x) \qquad \text{(E-3)}$$

where:

$\hat{y}_m(x)$ is the updated prediction at step m,

$\hat{y}_{m-1}(x)$ is the previous prediction,

$h_m(x)$is the prediction of the new tree fitted to the residual errors,

γ is the learning rate, controlling the contribution of each tree.

Gradient Boosting is particularly effective in capturing complex, non-linear relationships in hydrocarbon production data.

Support Vector Regression uses a margin-based approach to estimate continuous outputs. It seeks to find a function $f(x)$ that deviates from the actual target y_i by no more

than a predefined margin $\varepsilon\varepsilon$, and simultaneously remains as flat as possible. The optimization problem for SVR can be written as (Platt, 2000; Vapnik, 2000; Chang and Lin, 2022):

$$\min_{w,b} \left(\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} L_\epsilon \left(y_i, f(x_i) \right) \right) \qquad \text{(E-4)}$$

where:

w is the weight vector that defines the orientation of the hyperplane,

b is the bias term (or interecept) in the model,

C is a regularization parameter controlling the trade-off between maximizing the margin and minimizing the error,

L_ϵ is the loss function (epsilon-insensitive loss),

y_i is the actual value,

$f(x_i)$ is the predicted output for input (x_i).

SVR is particularly useful for its robustness in high-dimensional and sparse datasets, such as those found in petroleum applications.

Artificial Neural Networks and Recurrent Neural Networks are capable of modeling complex non-linear relationships and temporal dependencies. In a standard feedforward ANN like a Multi-Layer Perceptron. (Rosenblatt, 1958; Sarker, 2021). The forward pass of an MLP is mathematically defined as:

$$a^{(l)} = f\left(W^{(l)} a^{(l-1)} + b^{(l)} \right) \qquad \text{(E-6)}$$

where:

$a^{(l)}$ is the activation of the neurons in layer l,

$a^{(l-1)}$ is the activation of the neurons in previous layer $l-1$,

$W^{(l)}$ is the weight matrix connecting layer $l-1$ to layer l,

$b^{(l)}$ is the bias vector for layer l,

f is the activation function (e.g., ReLU, Sigmoid, Tanh).

For time-series forecasting, RNNs process sequences by incorporating previous states into the prediction of the current state:

$$h_t = f(W_h h_{t-1} + W_x x_t + b) \quad \text{(E-7)}$$

where:

h_t is the hidden state at time step t,

h_{t-1} is the hidden state from the previous time step,

W_h is the weight matrix for the hidden state,

W_x is the weight matrix for the input,

x_t is the input at time step t,

b is the bias vector,

f is the activation function (e.g., Tanh, ReLU).

This formulation enables RNNs to model time-dependent patterns in well performance and production decline curves (McCulloch and Pitts, 1943; Rosenblatt, 1958; Peter et al., 2019; Sarker, 2021).

Appendix F. Gadiant Boosting Algorithm

Table F1. Generalized boosting of decision trees (Friedman, 2002).

Step	Generalized Gradient Tree Boosting
1	$F_0(x) = arg\ min_\gamma \sum_{i=1}^{n} \Psi\,(y_i, \gamma)$
2	For m = 1 to M do:
3	$\tilde{y}_{im} = -\left[\frac{\delta \Psi_{(y_i, F(x_i))}}{\delta F(x_i)}\right]_{F(x)=F_{m-1}(x)}, i = 1, n$
4	$\{R_{lm}\}_1^L = L - terminal\ node\ tree(\{\tilde{y}_{im}, x_i\}_1^n)$
5	$\gamma_{lm} = argmin_\gamma \sum_{x_i \epsilon R_{lm}} \Psi(y_i, F_{m-1}(x_i) + \gamma)$
6	$F_m(x) = F_{m-1}(x) + \nu \cdot \gamma_{lm} 1(x\ \epsilon\ R_{lm})$
7	End For

Appendix G. SVR Mathematical Formulation

The SVR algorithm's uses $\varepsilon - SVR$ a classic SVR model aimed at finding a flat function, which has a small (ε) error from the obtained target (Moradzadeh et al., 2020). Consider a set of training points $\{(x_i, z_i),\cdots,(x_n, z_n)\}, where\ x_i \in \mathbb{R}^n$ is a feature vector and $z_i \in \mathbb{R}^1$is the target output; assuming parameters $C > 0\ and\ \epsilon > 0$, the standard form of $\varepsilon - SVR$ (Vapnik, 2000; Chang and Lin, 2022) becomes :

$$\underset{\omega, b, \xi, \xi^*}{min} \quad \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{n}\xi_i + C\sum_{i=1}^{n}\xi_i^* \qquad \text{(G-1)}$$

C is the penalty factor, ω is the coefficient of the function, b is the bias constant, n is the number of the last instance of the input data, ξ_i and ξ_i^* are the slack variables. Where b is the bias constant, subject to:

$$\begin{matrix} \omega^T\phi(x_i) + b - z_i & \leq \epsilon + \xi_i \\ z_i - \omega^T\phi(x_i) - b & \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq 0 \end{matrix}$$

$$i = 1,\cdots,n$$

The dual problem is:

$$\underset{\alpha, \alpha^*}{min} \quad \frac{1}{2}(\alpha - \alpha^*)^T Q(\alpha - \alpha^*) + \epsilon\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) + \sum_{i=1}^{n} z_i(\alpha_i - \alpha_i^*) \qquad \text{(G-2)}$$

Subject to:

$$e^T(\alpha - \alpha^*) = 0,$$

$$0 \leq \alpha_i, \alpha_i^* \leq C, i = 1,\cdots,n$$

where $Q_{ij} = K(x_i, x_j) = \phi(x_i)^T\phi(x_j)$ is the Kernel Function.

After solving Eq (X), the approximate function is:

$$\sum_{i=1}^{n}(-\alpha_i + \alpha_i^*)K(x_i, x) + b \qquad \text{(G-3)}$$

Wang et al. (2021) document the applicability of the RBF kernel to predict nonlinear regression problems. The RBF is highly flexible by adjusting the kernel function coefficient γ () shown as:

$$\mathrm{K}(x_i, x_j) = exp\left(-\gamma\left|x_i - x_j\right|^2\right), \gamma > 0 \qquad \text{(G-4)}$$

Where γ is the kernel function coefficient, C is the penalty factor, and ε is the maximum deviation.

Appendix H. Scaling Metrics

To ensure consistent feature scaling across input variables, this study employed both Min-Max Normalization and Z-Score Standardization—each chosen based on the distribution and nature of the data. These scaling techniques are essential for petroleum engineering datasets, where variable magnitudes can span multiple orders (e.g., permeability in microdarcies vs. pressure in psi). Without proper scaling, machine learning models, especially those that rely on distance metrics or gradient descent, can suffer from poor convergence and biased learning.

Min-Max Normalization was applied to features such as porosity and permeability to transform their values into a uniform range between 0 and 1. This scaling method is particularly useful for features with bounded domains and low variance, ensuring uniform contribution to model training:

$$dom(X) \rightarrow [0,1] \quad \text{(H-1)}$$

$$x_i' \leftarrow \frac{x_i - min_x}{max_x - min_x} \quad \text{(H-2)}$$

where:

x_i = initial value of a data point,

x_i' = normalized value of a data point x_i,

min_x = minimum value for feature x,

max_x = maximum value for feature x.

Z-Score Standardization was used for pressure and temperature—features expected to follow approximately normal distributions. This method centers the data around zero and scales it based on standard deviation, improving numerical stability during model training:

$$dom(X) \rightarrow \mathbb{R} \quad \text{(H-3)}$$

$$x_i' \leftarrow \frac{x_i - \bar{x}}{s} \quad \text{(H-4)}$$

where:

x_i = initial value of a data point,

$\bar{x}$ = mean of the feature x,

x_i' = standardized value of data point x_i,

s = standard deviation of the feature x in the dataset.

For features with skewed distributions or outliers, Robust Z-Score Standardization was implemented. This approach substitutes the mean and standard deviation with the median and interquartile range (IQR), providing greater resistance to extreme values:

$$dom(X) \rightarrow \mathbb{R} \quad \text{(H-5)}$$

$$x_i' \leftarrow \frac{x_i - \tilde{x}}{IQR_x} \quad \text{(H-6)}$$

where:

x_i = initial value of a data point,

$\tilde{x}$ = median of feature x in the dataset,

x_i' = robust-scaled value of data point x_i,

IQR_x = the interquartile range of the feature x.

These standardized and normalized inputs contribute to improved model interpretability, training convergence, and generalization in both flow pattern classification and production forecasting tasks.

Appendix I. Statistical Evaluation Metrics

To assess the performance of machine learning models used in this study for both classification and regression tasks, several statistical evaluation metrics were employed. These metrics provide insight into model accuracy, precision, robustness, and generalizability.

R-Squared (Coefficient of determination).

The coefficient of determination R^2 quantifies how well a regression model explains the variability of the response variable. It is defined as the proportion of the variance in the dependent variable that is predictable from the independent variables (Willmott and Matsuura, 2005). A value of $R^2 = 1$ indicates a perfect fit, while $R^2 = 0$ indicates that the model explains none of the variance.

Risidual Sum of Squares (RSS):

$$RSS = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \qquad \text{(I-1)}$$

Regression Sum of Squares (SS_R):

$$SS_R = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 \qquad \text{(I-2)}$$

Total Sum of Squares (SS_T):

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad \text{(I-3)}$$

The correlation coefficient (R^2) is defined as the ratio of the SS_R and SS_T. R^2 is measured on a scale $R^2 \in [0,1]$ where more variance is accounted for the close the value is to 1.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SS_R}{SS_T} = 1 - \frac{RSS}{SS_T} \qquad \text{(I-4)}$$

Where y_i is the observed dependent variable at i^{th} point, $\hat{y}_i$ is the predicted value of the dependent variable at i^{th} point, $\bar{y}$ is the mean value of the dependent variables (Willmott and Matsuura 2005).

Root-Mean-Squared Error.

The $RMSE$ represents how far, on average, the residuals are from zero (Kuhn and Johnson 2016) The residual errors are the difference between the observed values and predicted values. Minimizing these errors enhances the model's quality, leading to more accurate predictions.

$$RMSE = [n^{-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2]^{\frac{1}{2}} \qquad \text{(I-5)}$$

Where y_i is the observed dependent variable at i^{th} point, $\hat{y}_i$ is the predicted value of the dependent variable at i^{th} point, $\bar{y}$ is the mean value of the dependent variables.

Mean Squared Error.

MSE represents the average squared difference between actual and predicted values. It is similar to RMSE but retains the squared unit of the target variable, which may reduce interpretability:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \text{(I-6)}$$

where y_i represents the observed value at the ith data point, and $\hat{y}_i$ represents the predicted value at the same point.

Mean Absolute Error.

MAE measures the average magnitude of errors without considering their direction. It is less sensitive to outliers than $RMSE$ and is more interpretable since it uses the same units as the target variable:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \text{(I-7)}$$

where y_i represents the observed value at the ith data point, and $\hat{y}_i$ represents the predicted value at the same point.

Kappa Statistic (Cohen's Kappa).

Cohen's Kappa evaluates classification accuracy while accounting for agreement that could occur by chance. It is particularly useful when class distributions are imbalanced:

$$Kappa = k = \frac{p_o - p_e}{1 - p_e} \quad \text{(I-8)}$$

Where p_e is the probability agreement by chance and p_o is the observed agreement. A Kappa value of 1 indicates perfect agreement, while 0 indicates no better than chance.

Accuracy.

Accuracy is calculated as the ratio of true positives (TP) and true negatives (TN) to the total number of predictions made (Kuhn and Johnson, 2016):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{(I-9)}$$

True positives and true negatives represent correctly predicted cases, where TP refers to correctly identified positive instances and TN refers to correctly identified negative

instances. Conversely, false positives (FP) occur when a model incorrectly predicts a positive case, and false negatives (FN) occur when a model incorrectly predicts a negative case. While accuracy is intuitive, it can be misleading in imbalanced datasets, making metrics like Kappa essential for a more nuanced evaluation.

These metrics were chosen for their ability to capture both the accuracy and reliability of the machine learning models developed in this study, ensuring their suitability for real-world applications such as flow pattern classification and hydrocarbon production forecasting.

Appendix J. Mathematical Derivation of Fluid Flow Dimensionless Groups

For multiple phase flow regimes each of variables, $D, \Delta\rho g, v_s, \sigma, \mu, \rho$, each of these variables can be expressed dimensionally in terms of the primary dimensions of mass (M), length (L), and time (T). Derivation of Π Groups using Buckingham Π Theorem:

$$[D] = M^0 L T^0$$

$$[\Delta\rho g] = M L^{-2} T^{-2}$$

$$[v_s] = M^0 L T^{-1}$$

$$[\sigma] = M T^{-2}$$

$$[\mu] = M L^{-1} T^{-1}$$

$$[\rho] = M L^{-3}$$

The purpose of dimensional analysis is to determine the exponents of real numbers $x_1, x_2, x_3 \ldots, x_n$ such that the power production of these variables are zero. Mathematically, we have:

$$[D^{x_1}(\Delta\rho g)^{x_2} {v_s}^{x_3} \sigma^{x_4} \mu^{x_5} \rho^{x_6}] = M^{(x_2+x_4+x_5+x_6)} L^{(x_1-2x_2+x_3-x_5-3x_6)} T^{(-2x_2-x_3-2x_4-x_5)}$$

The power product will be dimensionless if and only if the following conditions are met:

$$x_2 + x_4 + x_5 + x_6 = 0$$

$$x_1 - 2x_2 + x_3 - x_5 - 3x_6 = 0$$

$$-2x_2 - x_3 - 2x_4 - x_5 = 0$$

The linear equation can be written in the following matrix format:

$$\begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & -2 & 1 & 0 & -1 & -3 \\ 0 & -2 & -1 & -2 & -1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Applying Gaussian-Jordan elimination methods yields:

$$\begin{bmatrix} 1 & 0 & 0 & 2 & 2 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & -1 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

We solve the equation system to yield the following results:

$$x_{1=} -2x_4 - 2x_5 - x_6$$

$$x_2 = -x_4 - x_5 - x_6$$

$$x_3 = x_5 + 2x_6$$

We supplement the above equations with the following additional solutions:

$$x_4 = x_4$$
$$x_5 = x_5$$
$$x_6 = x_6$$

They constitute the required solution to the dimensional analysis problem, can be rearranged as a linear combination of 3 vectors as follows:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} -2 \\ -1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} x_4 + \begin{bmatrix} -2 \\ -1 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} x_5 + \begin{bmatrix} -1 \\ -1 \\ 2 \\ 0 \\ 0 \\ 1 \end{bmatrix} x_6$$

As $x_4, x_5, and\ x_6$ can be any arbitrary number, we can choose $x_4 = -1\ and\ x_5 = x_6 = 0$ then we have the first solution.

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 0 \\ -1 \\ 0 \\ 0 \end{bmatrix}$$

$$\Pi_1 = \frac{gD^2\Delta\rho}{\sigma}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ -1 \\ 0 \\ -1 \\ 0 \end{bmatrix}$$

$$\Pi_2 = \frac{D^2\Delta\rho g}{v_{sj}\mu_j}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ -2 \\ 0 \\ 0 \\ -1 \end{bmatrix} x_6$$

$$\Pi_3 = \frac{\Delta\rho gD}{\rho_j v_{sj}^2}$$

Appendix K. Data Analysis Flow Patterns

This section presents the visualization and interpretation of dimensionless groups used to classify gas-liquid two-phase flow regimes. By plotting combinations of derived dimensionless parameters against observed flow patterns, the analysis evaluates the separability and clustering of flow regimes. These visualizations offer insight into how gravitational, buoyancy, and interfacial forces interact with pipe inclination and fluid properties, providing a basis for developing predictive machine learning models. The absence of smooth transitions observed in these scatter plots also suggests the discrete nature of flow regime classification, reinforcing the need for advanced classification algorithms.

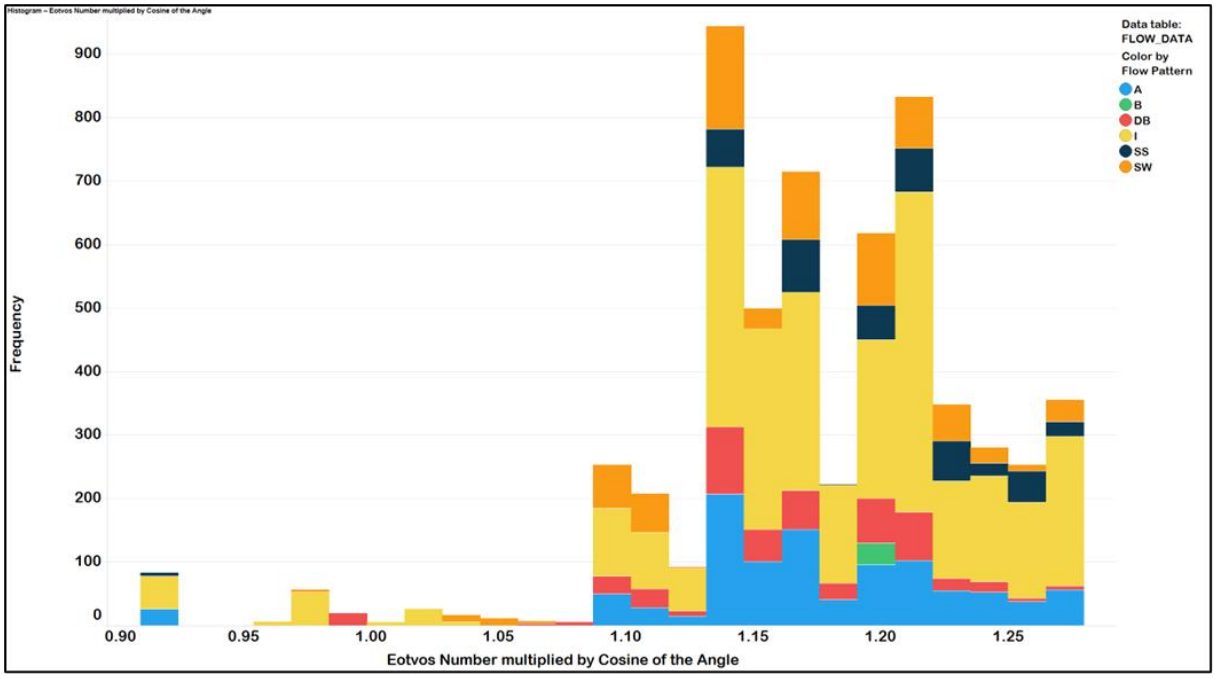

Figure K1. The flow pattern distribution for the $Eo^{0.25} * \cos \measuredangle_{pipe}$ highlights the distribution of flow patterns across different values of the Eo, capturing the influence of buoyancy and inclination on flow regime transitions. The clustering of specific flow patterns at distinct values supports its use as a predictive feature in flow modeling.

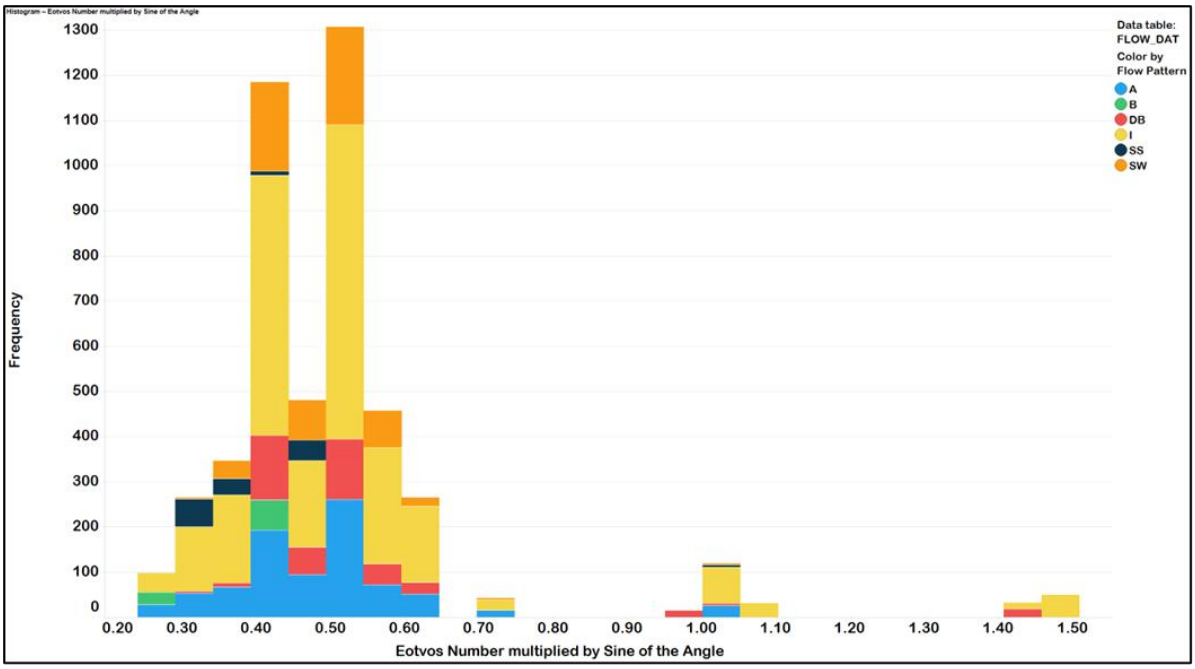

Figure K2. The flow pattern distribution for $Eo^{0.15} * \sin \measuredangle_{pipe}$ highlights the influence of gravitational and interfacial forces on flow regimes, showing distinct clustering of flow patterns at specific values. The distribution supports the relevance of this dimensionless parameter in predicting multiphase flow behavior.

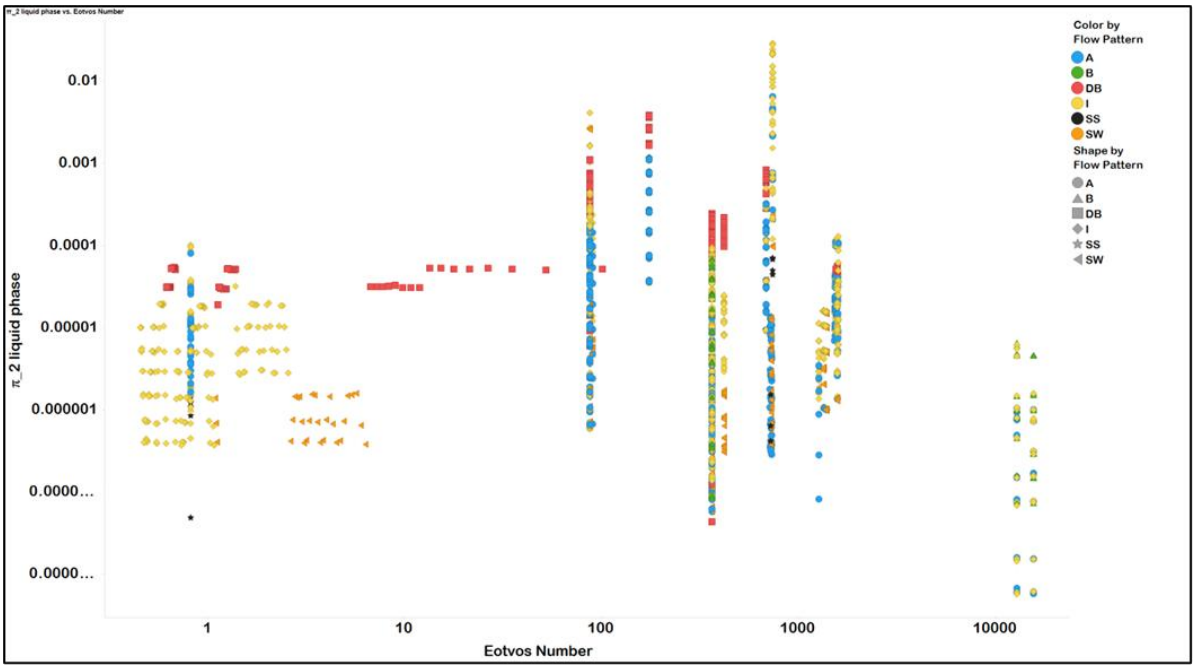

Figure K3. Plot of Π_2 and Π_1 with existing flow patterns in gas-liquid two-phase flow for all fluids in all pipes. Unlike traditional flow regime maps that show continuous phase transitions, the discrete clustering of flow patterns here suggests that these dimensionless groups act as strong separators of multiphase flow behavior. However, the lack of smooth transitions between patterns indicates that additional considerations may be needed to fully capture the complexity of real-world flow regimes.

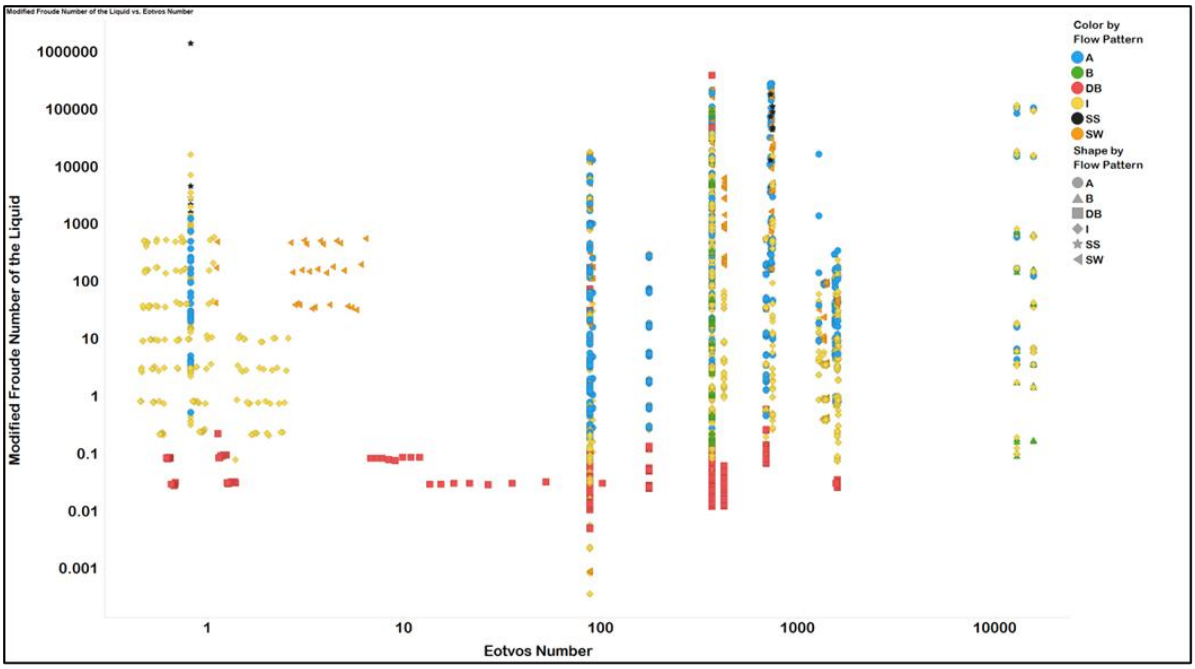

Figure K4. Plot of Π_3 and Π_1 with existing flow patterns in gas-liquid two-phase flow for all fluids in all pipes. The structured clustering highlights the ability of these dimensionless groups to differentiate multiphase flow behavior. However, the qualitative nature of the flow regime transitions may require further validation to ensure that gradual phase shifts are adequately represented.

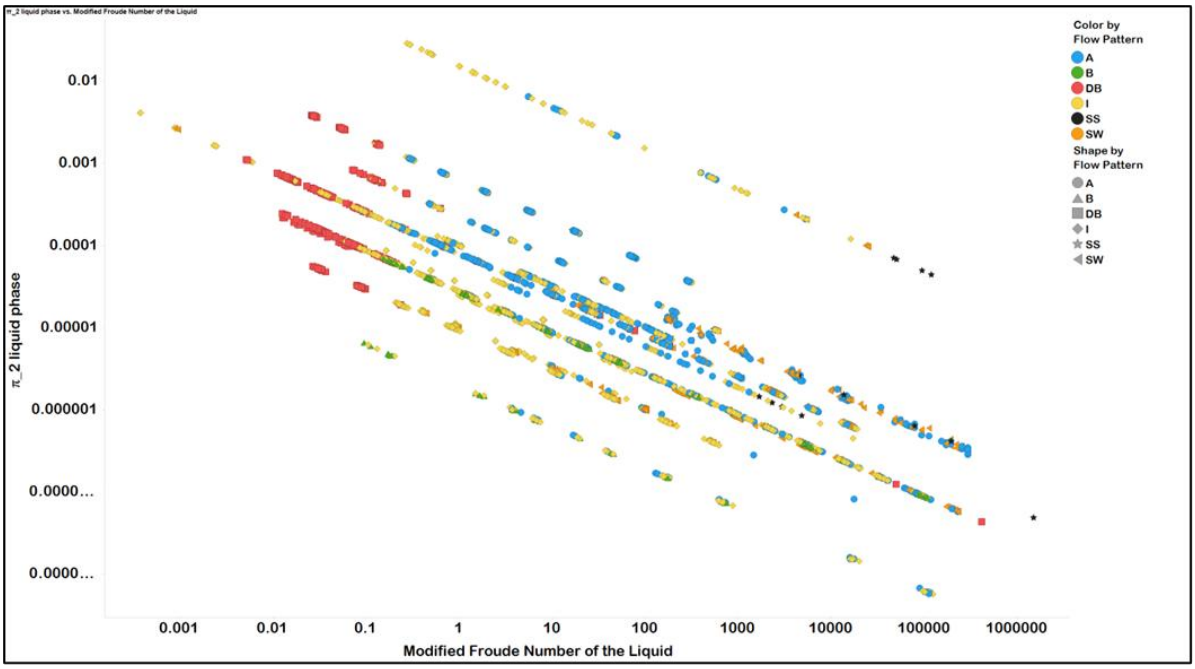

Figure K5. Plot of Π_3 and Π_2 with existing flow patterns in gas-liquid two-phase flow for all fluids in all pipes. The structured and linear clustering of flow patterns suggests a strong correlation between these dimensionless parameters, indicating their predictive capability in distinguishing flow regimes. However, unlike traditional flow regime maps, this visualization lacks smooth transitions between patterns, which may require additional validation or alternative visualization techniques to better represent gradual phase changes.

Appendix L. Reservoir Property Calculations

Accurate reservoir property estimation is essential for hydrocarbon production forecasting and machine learning model input preparation. This appendix outlines the formulas used to calculate key reservoir and fluid properties, including oil and gas density, viscosity, temperature, and pressure. These derived values were used to normalize inputs and support dimensionless group formation, contributing to consistent and physically meaningful model training.

The oil density is calculated using the oil API in the formula:

$$\rho_{oil} = \frac{141.5}{API+131.5} * \rho_{H_2O} \quad \text{(L-1)}$$

This formula relates oil density to API gravity, which provides a convenient empirical measure for light or heavy crude characterization.

In terms of specific gas gravity, the gas density is calculated from (Lee and Wattenbarger 1996):

$$\rho = \frac{p(28.963\gamma_g)}{10.732zT} = \frac{2.7p\gamma_g}{zT} \quad \text{(L-2)}$$

where:

ρ = gas density,

p = pressure,

γ_g = specific gas gravity (air = 1.0),

T = temperature,

Z = gas deviation factor.

The natural gas viscosity was determined by implementing Lee et al. (1966) method to correlate the mixture and pure component data simultaneously. This equation estimates gas density from pressure, specific gravity, temperature, and the gas deviation factor z. It is crucial for calculating gas-phase flow properties.

The gas viscosity proposed by Lee et al. (1966) is calculated in the form:

$$\mu = ke^{[X\rho^Y]} \quad \text{(L-3)}$$

where:

$$K = \frac{(9.4+0.02M)T^{1.5}}{209+19M+T} \quad \text{(L-4)}$$

$$X = 3.5 + \frac{986}{T} + 0.01M \quad \text{(L-5)}$$

$$Y = 2.4 - 0.2Y \quad \text{(L-6)}$$

Lee et al. (1966) model provides accurate gas viscosity estimates across a wide range of pressures and temperatures using molecular weight (M), temperature (T), and gas density (ρ).

The crude oil API is used to correlate the oil viscosity. Labedi (1992) proposed the following correlation for dead-oil viscosity of light crudes:

$$\mu_{od} = \frac{10^{9.224}}{(API)^{4.7013}(T_R)^{0.6739}} \quad \text{(L-7)}$$

where:

μ_{od} = dead-oil vicosity

API = stock-tank oil gravity @ atmosphere

T_R = reservoir temperature

This empirical correlation relates API gravity and reservoir temperature to dead oil viscosity, suitable for light crude oils under reservoir conditions.

The reservoir temperature is calculated as follows:

$$T_r = TVD * 0.0216 \quad \text{(L-8)}$$

where:

$TVD =$ total vertical depth, ft

This geothermal gradient approximation converts depth to temperature, assuming a regional gradient of $0.0216\ °F/ft$.

The reservoir pressure is calculated using hydrostatic gradient as follows:

$$P_i = 0.052 * TVD * MW \quad \text{(L-9)}$$

where:

$TVD =$ Total Vertical Depth,

$MW =$ mud weight

This formula estimates formation pressure from TVD and mud weight, assuming a pressure gradient of $0.052\ psi/ft\ per\ unit\ mud\ weight\ (ppg)$.

Appendix M. Dimensional Analysis of Reservoir Properties

The dimensionless reservoir characterization upscaling parameters derived from Buckingham Π Theorem are shown in a step-by-step procedure using algebraic theory. Consider that hydrocarbon production as a function of:

$$f(\mu, P, k, \rho, D, t), \quad \text{(M-1)}$$

Each of these variables can be expressed dimensionally in terms of the primary variables of mass (M), length (L), and time (T) as follows:

$$[\mu] = ML^{-1}T^{-1}$$

$$[P] = ML^{-1}T^{-2}$$

$$[k] = M^0L^2T^0$$

$$[\rho] = ML^{-3}T^0$$

$$[D] = M^0LT^0$$

$$[t] = M^0L^0T$$

Given $k = 3$ is the number of fundamental units (e.g., M, L, T) required in an absolute system for measuring the *n* = 6 kinds of quantity, the number of dimensionless products Π is:

$$i = n - k, \qquad \text{(M-2)}$$

$$i = 6 - 3 = 3\,, \qquad \text{(M-3)}$$

The purpose of DA is to determine the exponents of real numbers $x_1, x_2, x_3 \ldots, x_n$ such that the power production of these variables is zero. Mathematically:

$$[\mu^{x_1} P^{x_2} k^{x_3} \rho^{x_4} D^{x_5} t^{x_6}] = M^{(x_1+x_2+x_4)} L^{(-x_1-x_2+2x_3-3x_4+x_5)} T^{(-x_1-2x_2+x_6)}$$

The power product is dimensionless if and only if the following conditions are met:

$$x_1 + x_2 + x_4 = 0$$

$$-x_1 - x_2 + 2x_3 - 3x_4 + x_5 = 0$$

$$-x_1 - 2x_2 + x_6 = 0$$

The linear equation can be written in the following matrix format:

$$\begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ -1 & -1 & 2 & -3 & 1 & 0 \\ -1 & -2 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Applying Gaussian-Jordan elimination methods:

$$\begin{bmatrix} 1 & 0 & 0 & 2 & 0 & 1 \\ 0 & 1 & 0 & -1 & 0 & -1 \\ 0 & 0 & 1 & -1 & 0.5 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Solving the equation system to yielding:

$$x_{1=} -2x_4 - x_6$$

$$x_2 = x_4 + x_6$$

$$x_3 = x_4 - 0.5x_5$$

Supplementing the above equations with the following additional solutions:

$$x_4 = x_4$$

$$x_5 = x_5$$

$$x_6 = x_6$$

They constitute the required solution to the dimensional analysis problem, can be rearranged as a linear combination of 3 vectors as follows:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} -2 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} x_4 + \begin{bmatrix} 0 \\ 0 \\ 0.5 \\ 0 \\ 1 \\ 0 \end{bmatrix} x_5 + \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} x_6$$

As $x_4, x_5, and\ x_6$ can be any arbitrary number, choosing $x_4 = 1\ and\ x_5 = x_6 = 0$ yields:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} -2 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} x_4$$

$$\Pi_1 = \frac{kP\rho}{\mu^2} = \frac{L^2 \frac{M}{LT^2} \frac{M}{L^3}}{\frac{M^2}{L^2T^2}} = Dimensionless, \quad \text{(M-4)}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -0.5 \\ 0 \\ 1 \\ 0 \end{bmatrix} x_5$$

$$\Pi_2 = \frac{D}{k^{\frac{1}{2}}} = \frac{L}{(L^2)^{\frac{1}{2}}} = Dimensionless, \quad \text{(M-5)}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} x_6$$

$$\Pi_3 = \frac{Pt}{\mu} = \frac{\frac{M}{LT^2}T}{\frac{M}{LT}} = Dimensionless, \quad \text{(M-6)}$$

Appendix N. Test Train Results for EUR ML Model Integrated with DA.

The consistent R^2 values ranging from 0.5 to 0.8 for tests seen in **Figure N1** and 0.8 to 0.9 for train data seen in **Figure N2** across the CV scores highlight the generalization and robustness of the tree-based regression models. The models effectively capture the underlying patterns in the data. The ML models' strong predictive performance indicates they are reliable and suitable for practical applications.

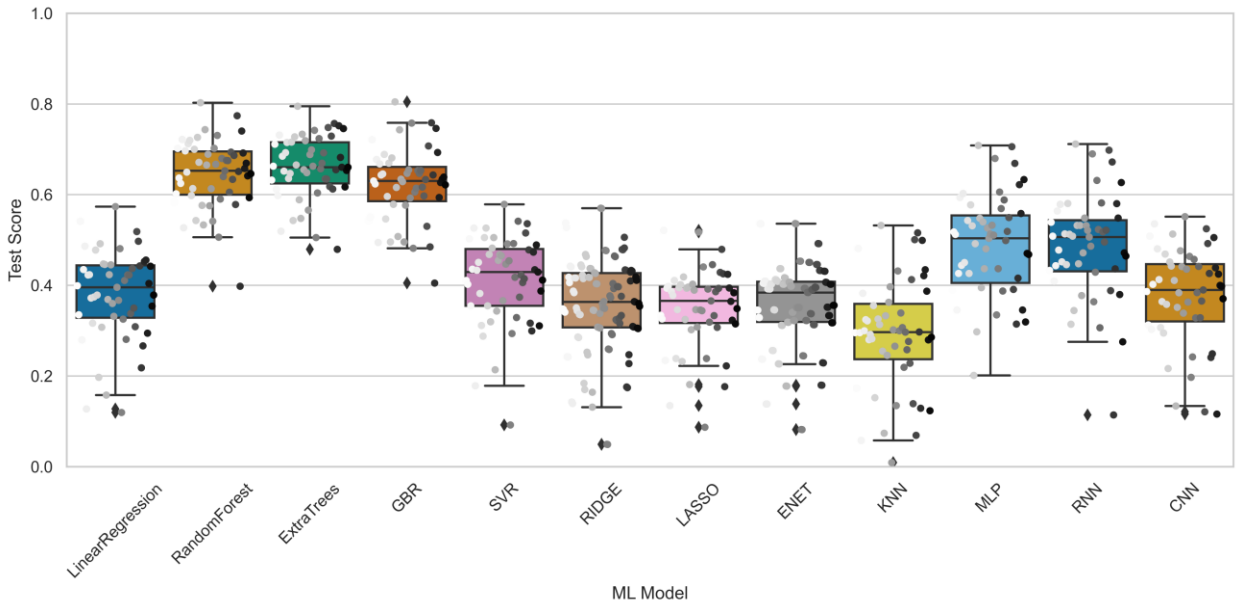

Figure N1. The performance of various machine learning models in predicting EUR, with tree-based models (RF, ET, and XGB) demonstrating consistent and strong predictive test performance (R^2>0.6) The relatively higher and stable scores of these models indicate their robustness in capturing underlying data patterns. Other models, including neural networks (MLP, RNN, CNN) and linear regressors, exhibit greater variability, highlighting differences in model generalization. The results suggest that tree-based models provide reliable predictions, making them well-suited for practical applications in hydrocarbon production forecasting.

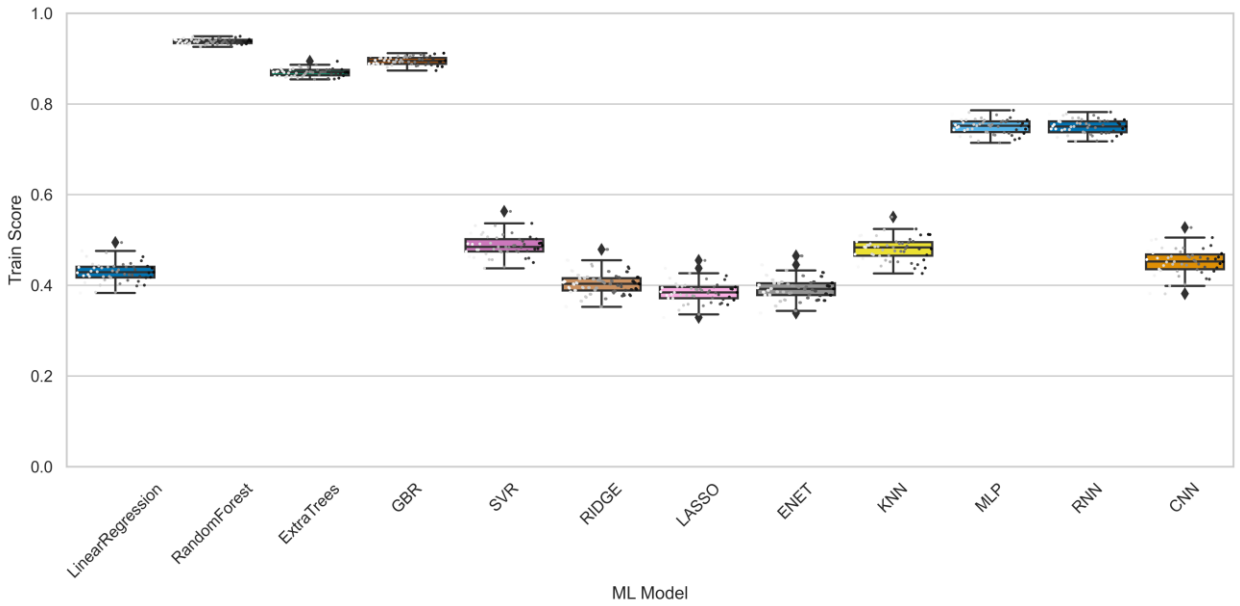

Figure N2. The training performance of various machine learning models, highlighting the exceptional generalization of tree-based models (RF, ET, and GBR), which achieve consistently high training scores (R^2>0.8). The high scores indicate that these models effectively capture underlying data patterns with minimal overfitting. In contrast, linear models and some neural network-based models (MLP, RNN, CNN) exhibit greater variability, suggesting differences in model complexity and adaptability to training data. The results reinforce the effectiveness of tree-based models for predictive accuracy in hydrocarbon production forecasting.

Appendix O. Scalar (EUR) Regression Model Hyperparameters.

The Random Forest Regression Model was selected for transfer from the scalar model to the vector model due to its strong predictive performance, robustness to overfitting, and ability to handle high-dimensional data. This model demonstrated superior accuracy in capturing complex production trends, making it well-suited for hydrocarbon production forecasting. The hyperparameters were optimized to balance bias-variance trade-off, ensuring both high generalization and computational efficiency. Table **O1** presents the final set of hyperparameters used in the Random Forest model, following the implementation guidelines from Pedregosa et al. (2011).

Table O1. Final static ML model hyperparameters (Pedregosa et al. 2011).

Parameter	Value	Description
Number of Estimators	50	The number of trees in the forest.
Criterion	Squared error	The function to measure the quality of a split.
Max Depth	50	The maximum depth of the tree.
Minimum Sample Splits	5	The minimum number of samples required to split an internal node.
Minimum Leaf Samples	3	The minimum number of samples required to be at a leaf node.
Minimum Weight Fraction	0.0	The minimum weighted fraction of the sum total of weights required to be at a leaf node.
Max Features	sqrt	The number of features to consider when looking for the best split.
Max Leaf Nodes	None	Grow trees with max leaf nodes in a best-first fashion.
Minimum Impurity Decrease	0.0	A node will be split if this split induces a decrease of the impurity >= to this value.
Bootstrap	False	Whether bootstrap samples are used when building trees.
Out-of-bag Score	False	Whether to use out-of-bag samples to estimate the generalization score.
Number of Jobs	None	The number of jobs to run in parallel.
Random State	None	Controls the randomness of the bootstrapping and feature sampling.
Verbose	0	Controls the verbosity during fitting and predicting.
Warm Start	False	Whether to reuse the previous fit and add more estimators.
CCP Alpha	0.0	Complexity parameter used for Minimal Cost-Complexity Pruning.
Max samples	None	The number of samples to draw from X to train each base estimator.
Monotonic State	None	Indicates the monotonicity constraint to enforce on each feature.

Appendix P. Temporal Feature Vector

The proposed temporal feature for the vector model to campure the series of time.

$$\overrightarrow{\Pi^{(\iota)}} = \left[\Pi_1^i, \Pi_2^i \ldots, \Pi_{n-1}^i, \Pi_{n_t}^i\right] \in \mathbb{R}^n$$

$$where:$$

- $n = (n-1) + 1 = number\ of\ static\ features + 1\ temporal\ feature$
- $\Pi_{n_t}^i\ represents\ the\ temporal\ aspect\ of\ \ sample\ i\ @\ time\ t$

$$\Pi' = \left[\begin{array}{cccc|c} \Pi_1^1 & \Pi_1^1 & \cdots & \Pi_{n-1}^1 & \Pi_{n_t}^1 \\ \Pi_1^2 & \Pi_2^2 & \cdots & \Pi_{n-1}^2 & \Pi_{n_t}^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \Pi_1^m & \Pi_1^m & \cdots & \Pi_{n-1}^m & \Pi_{n_t}^m \end{array}\right] \in \mathbb{R}^{m \times n}$$

Appendix Q. Typer Curve and MAE Bounds

A blind test was conducted using a publicly disclosed type curve from a reputable Permian Basin producer to evaluate the accuracy of the machine learning model. A set of parameters was generated based on the location and proppant intensity specified for the type curve. Dimensionless parameters for a hypothetical well were determined using nearby wells to test against the disclosed type curve.

The results, shown in **Appendix Q** (Figures **Q1** and **Q2**), demonstrate that the machine learning model, even with limited data, can generate a type curve that closely aligns with the producer's reference type curve. The predicted production follows the expected decline trends, confirming the model's ability to generalize well behavior. However, as training data becomes sparse, the predicted curve begins to flatten, indicating a reduction in predictive accuracy due to limited historical production data. Increasing the volume of training wells in later production time frames would improve the model's ability to capture long-term decline trends.

To further validate the model's accuracy, the MAE was used to define upper and lower production boundaries around the predicted values. Figures **Q1** and **Q2** illustrate these MAE-based confidence bounds, which serve as uncertainty margins for the predictions. More than 80% of predicted well forecasts fell within the expected ± MAE range, confirming that the model maintains acceptable accuracy across different production scenarios. Fewer than 20% of test wells fell outside these boundaries, reinforcing the model's robustness.

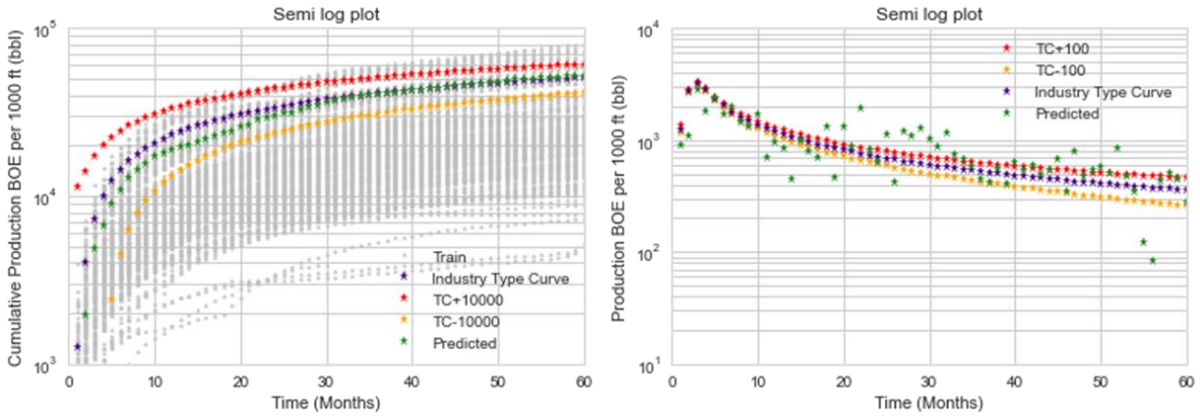

Figure Q1. The Mean Absolute Error bounds provide an uncertainty range, ensuring that more than 80% of the predicted forecasts fall within the $\pm MAE$ margin, reinforcing the model's reliability.

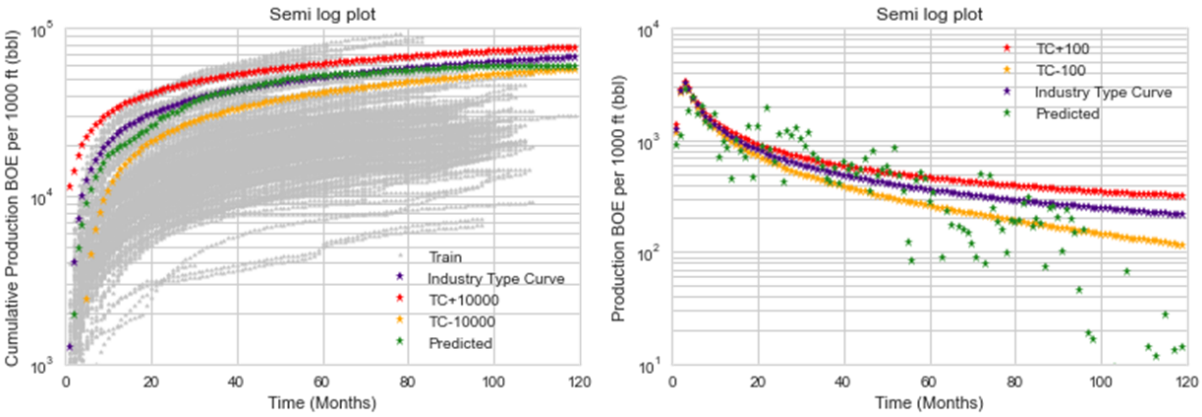

Figure Q2. The Mean Absolute Error provides an uncertainty range, reinforcing the need for additional historical data to enhance long-term predictive accuracy and reduce forecast variability.

The ability to generate type curves with bounded uncertainty using MAE provides a quantifiable measure of forecast confidence. This methodology is useful for operators assessing well performance and predicting future production trends with a data-driven, probabilistic approach.

Appendix R. Machine Learning Automated Dimensional Analysis.

Dimensional analysis plays a crucial role in simplifying complex engineering problems by reducing the number of independent variables into a set of dimensionless groups, preserving key system behaviors while significantly reducing computational and experimental costs. BuckinghamPy, a Python-based algorithm developed by Karam and Saad (2021), automates the derivation of Buckingham Π groups, streamlining the traditionally tedious process of dimensional analysis. By applying this algorithm, large-scale and otherwise infeasible experiments can be conducted efficiently while maintaining the fundamental physics governing reservoir and completion behavior.

Tables **R1** and **R2** present the possible dimensionless groups derived from completion and reservoir parameters, respectively. When using algorithm-generated Π groups, it is essential to verify their dimensionless nature to prevent computational errors and ensure the accuracy of machine learning applications in petroleum engineering applications.

Table R1. The possible combinations of completion related Π groups formulated using the Buckingham Π theorem. These groups help identify dominant physical interactions governing hydraulic fracturing performance, enabling more efficient and scalable production forecasting models.

Π_1	Π_2	Π_3
$\frac{P_T}{\sqrt[3]{V_s}g\rho}$	$\frac{S}{\sqrt[3]{V_s}}$	$\frac{Q_t\mu_f}{V_s^{\frac{4}{3}}g\rho}$
$\frac{\sqrt[3]{V_s}g\rho}{P_T}$	$\frac{S}{\sqrt[3]{V_s}}$	$\frac{Q_t\mu_f}{P_T V_s}$
$\frac{V_s^{\frac{4}{3}}g\rho}{Q_t\mu_f}$	$\frac{\sqrt[3]{V_s}}{S}$	$\frac{P_T V_s}{Q_t\mu_f}$
$\frac{g\rho S}{P_T}$	$\frac{V_s}{S^3}$	$\frac{Q_t\mu_f}{P_T S^3}$
$\frac{g\rho S^4}{Q_t\mu_f}$	$\frac{V_s}{S^3}$	$\frac{P_T S^3}{Q_t\mu_f}$
$\frac{\sqrt[3]{Q_T}\sqrt[3]{\mu_f}\, g\rho}{P_T^{\frac{4}{3}}}$	$\frac{P_T V_s}{Q_t\mu_f}$	$\frac{\sqrt[3]{P_T}\, S}{\sqrt[3]{Q_T}\sqrt[3]{\mu_f}}$
$\frac{V_s}{S^3}$	$\frac{P_T}{g\rho S}$	$\frac{Q_t\mu_f}{g\rho S^4}$
$\frac{V_s(g\rho)^3}{P_T^3}$	$\frac{g\rho S}{P_T}$	$\frac{Q_t\mu_f(g\rho)^3}{P_T^4}$
$\frac{V_s(g\rho)^{\frac{3}{4}}}{Q_T^{\frac{3}{4}}\mu_f^{\frac{3}{4}}}$	$\frac{\sqrt[4]{g\rho}\, S}{\sqrt[4]{Q_T}\sqrt[4]{\mu_f}}$	$\frac{P_T}{\sqrt[4]{Q_T}\sqrt[4]{\mu_f}(g\rho)^{\frac{3}{4}}}$

Table R2. The possible combinations of reservoir related Π groups. These dimensionless groups facilitate the development of generalized machine learning models, ensuring applicability across diverse reservoir conditions while preserving underlying flow physics.

Π_4	Π_5	Π_6
$\frac{kP_i\rho}{\mu^2}$	$\frac{D}{k^{\frac{1}{2}}}$	$\frac{P_i t}{\mu}$
$\frac{k\rho}{P_i t^2}$	$\frac{\mu}{P_i t}$	$\frac{D}{k^{\frac{1}{2}}}$
$\frac{k\rho}{\mu t}$	$\frac{P_i t}{\mu}$	$\frac{D}{k^{\frac{1}{2}}}$
$\frac{P_i D^2 \rho}{\mu^2}$	$\frac{k}{D^2}$	$\frac{P_i t}{\mu}$
$\frac{D^2\rho}{P_i\, t^2}$	$\frac{k}{D^2}$	$\frac{\mu}{P_i t}$
$\frac{D^2\rho}{\mu t}$	$\frac{k}{D^2}$	$\frac{P_i t}{\mu}$
$\frac{kP_i\rho}{\mu^2}$	$\frac{\sqrt{P_i} D\sqrt{\rho}}{\mu}$	$\frac{P_i t}{\mu}$
$\frac{k}{D^2}$	$\frac{\mu}{\sqrt{P_i} D\sqrt{\rho}}$	$\frac{\sqrt{P_i} t}{D\sqrt{\rho}}$
$\frac{k\rho}{P_i t^2}$	$\frac{\mu}{P_i t}$	$\frac{D\sqrt{\rho}}{\sqrt{P_i} t}$
$\frac{k}{D^2}$	$\frac{P_i D^2 \rho}{\mu^2}$	$\frac{\mu t}{D^2\rho}$
$\frac{k\rho}{P_i t}$	$\frac{P_i t}{\mu}$	$\frac{D\sqrt{\rho}}{\sqrt{\mu}\sqrt{t}}$
$\frac{k}{D^2}$	$\frac{P_i t^2}{D^2\rho}$	$\frac{\mu t}{D^2\rho}$
$\frac{kP_i\rho}{\mu^2}$	$\frac{D}{k^{\frac{1}{2}}}$	$\frac{\mu t}{k\rho}$
$\frac{P_i t^2}{k\rho}$	$\frac{\mu t}{k\rho}$	$\frac{D}{k^{\frac{1}{2}}}$
$\frac{\mu}{\sqrt{k}\sqrt{P_i}\sqrt{\rho}}$	$\frac{D}{k^{\frac{1}{2}}}$	$\frac{\sqrt{P_i} t}{\sqrt{k}\sqrt{\rho}}$