UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

DATA-DRIVEN MODELS FOR BOTTOMHOLE PRESSURE PREDICTION: APPLICATIONS IN HYDRAULIC FRACTURING, GAS LIFT, AND NITROGEN-LIFTED WELLS

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

SAMUEL NASHED
Norman, Oklahoma
2026

DATA-DRIVEN MODELS FOR BOTTOMHOLE PRESSURE PREDICTION: APPLICATIONS IN HYDRAULIC FRACTURING, GAS LIFT, AND NITROGEN-LIFTED WELLS

A THESIS APPROVED FOR THE
MEWBOURNE SCHOOL OF PETROLEUM AND GEOLOGICAL ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Rouzbeh Ghanbar Moghanloo, Chair

Dr. Raki Sahai

Dr. Hamidreza Karami

ACKNOWLEDGEMENTS

I want to thank my mentor and advisor, Dr. Rouzbeh Moghanloo, who has provided guidance, support, and confidence in my capabilities during my master's degree studies at the University of Oklahoma. I am truly grateful to have a chance to become his teaching assistant, and it thoughtfully enriched my academic experience and helped me to advance my professional and research abilities. He has played a key role in this work through his mentorship, wisdom, and support, as well as in assisting me to reach a significant milestone in my academic and professional path.

I also have the privilege to acknowledge my thesis committee members, Dr. Raki Sahai and Dr. Hamidreza Karami, whose time, useful remarks, and insightful review of this research were valuable.

I would like to thank the faculty and the staff of the Mewbourne School of Petroleum and Geological Engineering that provided academic and administrative support. I would also like to specifically thank Kathleen Shapiro, who has been of great help, guidance, and clarification during this process.

Lastly, I would like to say a sincere thank you to my wife, Nermeen, and her constant support, patience, and encouragement in the hard and fulfilling time of this endeavor. I am also grateful to my family, friends, and loved ones who provided unremitting support that enabled me to achieve this goal.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

Abstract

Precise forecasting of bottomhole pressure is the heart of optimizing the performance of wells, their operational efficiency, and decision-making in several petroleum engineering applications. Traditional physics-based and empirical correlations are frequently not accurate in complex conditions of multiphase flow, whereas direct downhole measurements are also costly and operationally constrained. This research developed individual machine learning (ML) models for each application to predict bottomhole pressure in hydraulic fracturing, gas lift wells, and nitrogen-lifted wells.

Large field data were gathered from vertical wells, consisting of 42 hydraulically fractured wells, 304 gas-lift wells, and 518 nitrogen-lifted wells. A scientific approach was used and included ranking of the features, preprocessing of data, and a series of ML models, such as linear regression, tree-based algorithms, ensemble models, neural networks, and symbolic regression. The standard statistical measures were used to assess model performance and were tested on cross-validation and blind tests on independent wells. Moreover, in the case of nitrogen-lifted wells, the results were compared to five empirical correlations that had been defined.

These findings indicate that the accuracy and generalization of the advanced ML models are better than traditional approaches. Symbolic regression also has interpretable equations, which are more transparent. The suggested framework will facilitate cost-efficient, scalable, and real-time monitoring of BHP without downhole gauges, which will facilitate optimization and intelligent field operations.

Chapter 1: Introduction

1.1 Background and Motivation

1.1.1 Importance of Bottomhole Pressure in Oil and Gas Operations

Bottomhole pressure (BHP) is one of the most important parameters in petroleum engineering, and it directly determines the optimization of production, reservoir management, and operational safety of various well interventions and production procedures. The correct BHP knowledge can contribute to the optimization of production rates, avoid the damage of formation, and make rational decisions concerning the artificial lift needs and well stimulation methods. BHP monitoring in hydraulic fracturing operations is critical to real-time treatment diagnostics and making critical decisions related to pumping schedules, proppant placement strategies, and early identification of operational problems like premature screenouts. The pressure data is essential in understanding the fracture geometry and assessing the treatment effectiveness after the end of the job. Likewise, in gas lift systems, accurate BHP specification allows the gas injection rate to be properly controlled to get maximum production with minimum gas usage. With nitrogen-lifted wells, a precise BHP forecast at coiled tubing depth is essential in determining the ideal nitrogen injection rates and volumes, modification of coiled tubing depth and velocity, and avoiding any complication of operations such as excessive pressure drawdown, proppant flowback, or intermittent flow.

1.1.2 Challenges in Real-Time Pressure Monitoring

Although it is highly essential, real-time BHP monitoring has significant technical and financial difficulties during field operations. Direct measurement involves the deployment of downhole pressure gauges, either by wireline or coiled tubing or completion tubing, and is a major operational task, with all costs being within the range of $50,000 to $150,000 per installation. The operation of these gauge deployments brings in complexity of operations, possible safety risks, and logistical limitations, which restrict their common usage at large populations of wells. Moreover, the issues of gauge reliability, calibration, and the nature of installing gauges in severe environments of the downhole only increase the challenges. The other method of determining BHP by surface measurements through physics-based models or empirical relationships has inherent inaccuracies, mostly in a system operating dynamically with complex multiphase flow, non-Newtonian rheology, and time-dependent fluid characteristics. There are also large calibration requirements of these computational methods to field data of particular fluid type, proppant characteristics, and wellbore geometry to enable generalizability and real-time applications.

1.1.3 The Role of Artificial Lift and Well Stimulation Technologies

Increasingly, modern petroleum recovery depends on artificial lift methods and well stimulation techniques to sustain economic rates of production as pressure drops in reservoirs or in unconventional reservoirs that need stimulation to be economically viable. Hydraulic fracturing is now essential in the production of tight formations and unconventional reservoirs, and the success of treatment with the desired geometry of the fractures and avoiding operational failures is largely reliant on the maintenance of the correct bottomhole pressure. Gas lift systems are currently among the most common systems of artificial lift deployed on a global scale, where gas is injected into the wellbore to decrease the pressure at the bottom of the flowing fluid and allow the reservoir fluids to be brought to the surface. Coiled tubing operations of nitrogen lifting offer temporary artificial lift solutions to well unloading and production testing, and precise pressure control is necessary to achieve performance optimization. All these applications require proper BHP forecasting to enhance the efficiency of its operations, safety, and production results.

1.2 Problem Statement

1.2.1 Limitations of Traditional BHP Prediction Methods

There are three main categories of conventional BHP prediction methodologies: physics-based mechanistic models, empirical correlations, and mixed methods of the two. Mechanistic multiphase flow models provide quite detailed, rigorous models of fluid dynamics in wellbores, which are based on the basic conservation equations but require large amounts of computational resources and large amounts of input data and must be heavily calibrated, which makes them impractical to operate in real-time operational settings. Existing empirical correlations based on experimental data and field observations, such as the Hagedorn-Brown, Beggs-Brill, Orkiszewski, Duns-Ros, and Fancher-Brown approaches, offer computationally fast alternatives but have low extrapolation. These correlations have been shown to return large inaccurate predictions when used on operational conditions, fluid systems, or well configurations other than those used to develop the correlations, resulting in large errors in predictions in field applications. Hybrid models seek to combine the benefits of either method but add a level of complexity to implementation and may need a lot of validation to guarantee that they can be trusted to operate under a wide variety of conditions.

1.2.2 Accuracy and Reliability Issues in Field Applications

Field applications in hydraulic fracturing, gas lift, and nitrogen lifting processes demonstrate that there are still problems with accuracy in using conventional prediction techniques. In hydraulic fracturing, the rheological behavior of the cross-linking fluids, especially those based on borate and other cross-linking systems, is multidimensional and is not represented well by conventional models, especially those based on standard correlations. Simple correction factors do not capture the effect of proppant on the friction pressure, particularly at high concentrations where complex multiphase flow and non-uniform flow behavior occur in vertical wellbores. Mechanistic models used in gas lift wells are difficult to calibrate and computationally expensive in real-time, whereas empirical correlations have been shown to be less precise over the wide variety of gas-liquid ratios, water cuts, and injection rates that are encountered in field operation. The presence of nitrogen-lifted wells creates especially difficult prediction environments, with transient multiphase flow regimes and a multiphase flow path of annular flow between the coiled tubing outer diameter and the inner diameter of the tubing.

1.2.3 Economic and Operational Constraints of Downhole Pressure Gauges

The direct method of measuring the pressure in the downhole with pressure gauges that are likely to give the most accurate data on BHP has extreme economic and operational limitations, which restricts their use on a large scale. The cost of gauge deployment, such as equipment rent, wireline or coiled tubing, rig time, and manpower, normally incurs over 50000 dollars per installation and up to 150000 dollars in the case of deep or difficult wells. These high-cost prices become excessively high to conduct regular monitoring of large well populations, especially in mature fields that have hundreds or thousands of wells to monitor. Others that are constraining operations are the complexity of deployment in deviated or horizontal wells, reliability issues during harsh downhole conditions, scarcity of gauge inventory during times of peak activity, or safety considerations related to wellbore access activities. Moreover, gauges can offer measurements only at discrete deployment intervals as opposed to continuous monitoring, so they are not as useful as real-time optimization of operations and dynamic decision-making in critical operations, e.g., hydraulic fracturing treatments or adjustment of artificial lift systems.

1.3 Research Objectives

1.3.1 Primary Objectives

The proposed study will create, test, and showcase superior machine learning models to accurately predict bottomhole pressure in real-time and through three important applications of petroleum engineering: hydraulic fracturing, gas lift wells, and nitrogen-lifted wells. The primary objectives include:

1. Developing predictive data-driven models, which are more accurate than conventional physics-based and empirical approaches.

2. Elimination of operational reliance on high-cost downhole pressure gauges by prediction using only easily available surface measurements.

3. Allows pressure monitoring solutions to be scaled and made cost-effective enough to be deployed on large populations of wells.

4. Offering interpretable alternatives based on symbolic regression to trade predictive power with physical transparency and engineering understanding.

1.3.2 Specific Aims for Each Application Domain

For hydraulic fracturing, the study develops models to predict bottomhole pressure at the perforation depth during cross-linked fluid injection, using inputs such as surface pressure, slurry flow rate, proppant concentration, tubing specifications, gauge depth, gel loading, and proppant properties. The model predictions are then validated against downhole gauge measurements from 42 vertical wells. The purpose of the gas lift application is to forecast flowing bottomhole pressure at perforation depth, based on completion parameters, reservoir properties and operating variables such as gas injection conditions, and rigorously tested through blind testing on 30 independent wells to establish generalizability. In the case of nitrogen-lifted wells, the aim includes predicting the bottomhole pressure at coiled tubing depth during lifting operations, providing performance benchmarking of five traditional empirical correlations (Hagedorn-Brown, Beggs-Brill, Orkiszewski, Fancher-Brown, and Duns-Ros), and proving the superiority by blind testing on 29 independent wells representing a variety of operational conditions.

1.4 Significance of the Study

1.4.1 Advancements in Predictive Modeling

This study will be the first comparative study on advanced machine learning algorithms applied specifically in bottomhole pressure prediction among various petroleum engineering tasks using vast field data. This study, as opposed to the past studies concentrated on single algorithms or synthetic data, offers systematic benchmarking of 10-16 models per application, including linear regression, tree-based, ensemble, support vector machines, neural networks, and symbolic regression, and runs them under the same conditions and the same validation methods. Integration of SHAP (SHapley Additive exPlanations) analysis offers an unprecedented understanding of feature significance and model decision-making methods, whereas genetic programming-based symbolic regression answers the long-standing critique of machine learning as opaque black-box technology by extracting explicit, interpretable mathematical equations with competitive predictive efficiency.

1.4.2 Practical Implications for Field Operations

The invented models allow bottomhole pressure monitoring in real-time, without downhole instrumentation, which allows the dynamic optimization of operational parameters through instantaneous feedback of pressure. Engineers can regulate the rates of gel injected in hydraulic fracturing, proppant concentrations, and pumping schedules to optimize fracture geometry and avoid screenouts. Gas lift operations have the advantage of optimum injection strategies that would maximize production and minimize the use of gas and eliminate the possibility of liquid loading or valve failures. The nitrogen lifting processes get improved efficiency due to the possibility of dynamic adjustment of the injection rates and the depth of coiled tubing to best maximize the production without having to face any complications in the operation. The computational performance of the deployed models makes it straightforward to integrate them with the current SCADA systems to provide real-time decision support as opposed to post-treatment analysis.

1.4.3 Cost-Effectiveness and Real-Time Monitoring Capabilities

The machine learning solution will provide order-of-magnitude cost savings versus conventional downhole gauge deployments and will only require parameters on the surface, which are collected using standard field equipment. The potential financial saving of up to 50,000-150,000 per well to deploy the gauge, in addition to keeping the prediction accuracy, results in large economic gains, especially when the operators have to operate with large well populations that need constant monitoring. Scalability of the method, whereby minimal computational resources are needed to do prediction inference even when significant initial training costs are required, makes it possible to deploy the method across hundreds or thousands of wells in parallel without the cost increasing proportionally. This economic and high-quality accuracy, which has been proven by intense field testing, makes machine learning models a breakthrough technology in monitoring bottomhole pressure in the contemporary petroleum industry.

1.5 Scope and Limitations

1.5.1 Geographical and Operational Context

This study relies solely on field data of vertical wells that have been used in the Western Desert of Egypt, and this includes certain geological formations, fluid systems, and practices that have been used in this area. The hydraulic fracturing data set deals with borate cross-linked guar-based fluids containing high-strength and intermediate-strength proppants. Gas lift wells are designed following the conventional vertical completion scheme with standardized gas injection systems. Nitrogen-lifted wells involve the use of coiled tubing operations that utilize certain equipment dimensions and nitrogen injection parameters. Although this narrow scope assures the data consistency and the possibility of rigid model development, it limits immediately the generalizability of the results to a much different geographical location, geological formation, or working practice without further validation.

1.5.2 Data Availability and Quality Considerations

The research data includes 42 hydraulic fracturing wells, 304 gas lift wells, and 518 nitrogen-lifted wells, which is a large but spatially concentrated amount of data collection. The quality of training data, its accuracy, and its representativeness are core elements of model performance, and the reliability of prediction relies on the fact that the field measurements are adequate and the sensors are calibrated. Missing values, outliers, and measurement errors are handled by data preprocessing procedures. The time range of data collection and the lack of some operational extremes or odd setups in the training data are limitations on model use.

1.5.3 Model Generalizability Constraints

Although rigorous validation using k-fold cross-validation, repeated random sampling, and blind dataset testing has proved the good performance of the models in the parameter range of training and validation sets, it is not clear how well the models will perform under conditions very different as compared to those used in training datasets. Models that are being tested with extreme parameter values or with new fluid systems or unusual completion designs can be subject to model validation or retraining before operational use. The narrow scope on vertical wells requires subsequent research before the transfer of developed models to horizontal or highly deviated wellbores, where the flow mechanics of the wells vary significantly. Also, the black-box characteristics of the best-performing neural networks can be faced with opposition in regulatory settings demanding physically transparent models, but symbolic regression approximations can be used to a certain extent to mitigate this problem.

1.6 Thesis Organization

This thesis will use five chapters to provide a thorough exploration of machine learning solutions to bottomhole pressure prediction in hydraulic fracturing, gas lift, and nitrogen-lifted well operations.

Chapter 1: Introduction provides the background and motive for the creation of machine learning-based BHP predictive models, the problem statement that reveals the weaknesses inherent in the previous method, the specific goals of the research in each application area, and the outline of the research meaning, extent, and limits.

Chapter 2: Literature Review contains a critical review of the available knowledge in the field, including the basic principles of bottomhole pressure in various well types; the classical prediction models based on the physics-based model; the correlation model; and the hybrid approach in the context of hydraulic fracturing, gas lift, and nitrogen lifting. It also includes the evolution of the machine learning method in petroleum engineering and its application in prediction of pressure.

Chapter 3: Methodology outlines the systematic research design that was used in this study, such as thorough data collection of the 42 hydraulic fracturing wells, 304 gas lift wells, and 518 nitrogen-lifted wells; detailed data preprocessing methodology (including quality assessment, normalization, and partition strategy); feature ranking analysis using correlation methodologies;

and systematic construction of various machine learning models, including linear methods, tree-based algorithms, ensemble techniques, neural networks, and symbolic regression.

Chapter 4: Results and Discussion provides a thorough assessment of the model performance in all three applications, comparing prediction accuracy by the use of common metric standards; extensive validation by use of k-fold cross-validation, repeated random sampling, and blind dataset validation; demonstration of model performance using real operational field data; quantitative benchmarking to traditional empirical correlations; and critical assessment of model limitations and constraints.

Chapter 5: Conclusions summarizes the research findings, outlining the significant results in each of the application areas, describing the new knowledge contributions to predictive modeling and petroleum engineering, and describing the practical implications of the research in the operations of the industry, such as cost reduction and the capability to make decisions in real-time; limitations in the study; and recommendations regarding future research directions to enhance the technology of data-driven bottomhole pressure prediction.

Chapter 2. LITERATURE REVIEW

2.1 Fundamentals of Bottomhole Pressure

2.1.1 Physical Principles and Governing Equations

Bottomhole pressure represents the total pressure exerted at a specific depth within a wellbore, comprising contributions from multiple physical phenomena governing fluid behavior in subsurface conditions[1]. The fundamental equation describing static bottomhole pressure in a vertical well follows from hydrostatic principles, as given by Equation (1).

$$P_{BH} = P_{\text{surface}} + \rho g h \tag{1}$$

where P_{BH} denotes bottomhole pressure, $P_{surface}$ represents surface pressure, ρ represents fluid density, g indicates gravitational acceleration, and h represents vertical depth. However, this simplified formulation applies only to static conditions with single-phase, incompressible fluids. Real petroleum engineering operations involve dynamic, multiphase flow conditions requiring substantially more complex mathematical representations[2].

During fluid flow in wellbores, the total pressure drop comprises three primary components: hydrostatic pressure, friction pressure, and acceleration pressure. The general pressure gradient equation for flowing conditions can be expressed as given by Equation (2).

$$\frac{dP}{dL} = \left(\frac{dP}{dL}\right)_{\text{hydrostatic}} + \left(\frac{dP}{dL}\right)_{\text{friction}} + \left(\frac{dP}{dL}\right)_{\text{acceleration}} \tag{2}$$

The hydrostatic component depends on fluid mixture density, which varies with phase composition, pressure, and temperature. For multiphase flow, calculating effective mixture density requires knowledge of liquid holdup, the fraction of pipe cross-sectional area occupied by the liquid phase[3]. The friction component arises from viscous shear stresses between flowing fluids and pipe walls, as well as interfacial friction between phases. Acceleration pressure becomes significant when fluid velocity changes substantially, particularly in wet gas wells or during rapid flow rate variations.

For non-Newtonian fluids commonly used in hydraulic fracturing operations, the rheological behavior is often described by the power-law model, which is characterized by the consistency index and the flow behavior index[4]. The apparent viscosity varies with shear rate according to Equation (3).

$$\mu_{\text{apparent}} = k'(\dot{\gamma})^{n'-1} \tag{3}$$

Cross-linked fracturing fluids exhibit particularly complex rheology, with viscosity dependent on shear history, temperature, and time since mixing. The cross-linking process transforms linear polymer solutions into three-dimensional gel networks, dramatically increasing viscosity and creating elastic properties that enhance proppant suspension capabilities[5].

The presence of proppant in fracturing slurries introduces additional complexity through particle-fluid interactions, settling effects, and non-homogeneous flow patterns. The effective slurry density exceeds base fluid density by an amount proportional to proppant concentration and specific gravity, directly impacting hydrostatic pressure. Friction pressure increases non-linearly with proppant concentration due to increased fluid viscosity, increased wall roughness effects from particle interactions, and potential flow regime transitions from homogeneous suspension to heterogeneous transport[6].

2.1.2 Factors Affecting BHP in Different Well Configurations

The dynamics of bottomhole pressure differ significantly with different well completions, operational modes, and fluid systems. The key factors in vertical wells are the depth of the wellbore, fluid density profile, flow rate, internal diameter of tubing, and fluid rheology[7]. The temperature variations in a wellbore affect the properties of fluids, in particular, gas density and viscosity, generating pressure-temperature coupling, which cannot be ignored in true prediction.

The horizontal and deviated wells bring in the new complexity of inclination angle effects on the multiphase flow patterns. The shift in vertical to horizontal orientation is a dramatic change in the effects of gravity on phase distribution and liquid hold-up. Stratified flow patterns in horizontal sections are known to have varying pressure gradients with those of bubble, slug, or annular flow regimes in vertical wellbores[8].

The bottomhole pressure during hydraulic fracturing processes critically depends on the ongoing change of the fluid properties during treatment. The sequence of pad stages with clean crosslinked gel to slurry stages with proppant forms time-varying profiles of rheology and density. Surface pressure, slurry injection rate, and proppant concentration are dynamic and demand real-time prediction capabilities of pressure. The kinetics of cross-linking of borate systems involves time-dependent viscosity growth, which complicates the use of static correlation.

Gas lift systems represent special pressure dynamics under the control of gas injection strategy. The depth of the injection point, the rate of gas injection, the injection operating pressure, and the design of production tubing all determine the multiphase flow regime and related pressure gradients. The ratio of gas to liquid is varying along the wellbore as the injected gas is expanding under lower pressure, producing spatially different flow patterns. Water cut also has an effect on the density of the mixture and changes in flow regime, especially the critical rate of gas injection needed to begin and maintain stable lift[9].

Nitrogen-lifted wells under coiled tubing lift behavior are affected by annular geometry that exists between the coiled tubing outer and production tubing inner diameters. The rate of nitrogen injection, depth of coiled tubing, and rate of production at the surface interplay to produce bottomhole pressure at the coiled tubing exit. The intermittent character of the nitrogen lifting processes, especially when unloading wells, results in time-resolved pressure profiles that are significantly different than steady-state gas lift conditions[10].

2.1.3 Measurement Techniques and Instrumentation

Direct bottomhole pressure measurement is mainly based on electronic pressure gauges that are sent down through wireline, slickline, or coiled tubing. Memory gauges are used to collect pressure and temperature data at specified sampling rates, which normally vary between seconds and minutes, based on the needs of the application. These are independent devices that record measurements within themselves, which are later recalled and analyzed when the gauge recovers. Surface-readout systems provide real-time data, allowing continuous monitoring of the operations, but they are more complex to deploy and more expensive.

Permanent downhole gauges fitted during well completion mean the ability to continuously monitor the pressure of wells without intervention operations. Such systems are especially useful in the critical wells where continuous surveillance justifies the heavy investment of capital. There are, however, gauge reliability issues, memory systems have a constraint on their battery life, and surface-readout configurations have a risk of failure, which precludes widespread adoption[11]. The extreme downhole condition, which is associated with high temperatures, corrosive fluids, mechanical vibration, and pressure cycling, has a detrimental effect on the long-term gauge performance[12].

The accuracy of a pressure gauge is determined by the quality of calibration, sensor drift properties, and environmental compensation. Quartz crystal gauges are of high precision with measurement uncertainties of less than 0.01 percent of full scale, but they are priced at a premium. Sensors based on strain gauges have reasonable accuracy (typically 0.1-0.25% full scale) at reduced cost and are predominant in the field[13]. Temperature compensation is very important to ensure accuracy throughout the thermal gradient experienced in wellbores, especially in deep or geothermally heated formations.

The aspect of deployment has a great impact on measurement quality and operational feasibility. Well control equipment, rig-up, and specialized expertise are needed in the wireline operations. The coiled tubing deployments are beneficial in the deviated or horizontal wells but add complexity and cost to the operations[14,15]. The gauge depth must be chosen reasonably to ensure that the pressure information is relevant and, at the same time, there are no restrictions, tight places, or working hazards. The second method of estimating bottom hole pressure using surface measurements does not present any deployment problems but introduces uncertainty in the model.

2.2 Bottomhole Pressure in Well Operations

2.2.1 Hydraulic Fracturing Operations

2.2.1.1 Overview of Hydraulic Fracturing

Hydraulic fracturing is the most commonly used well stimulation method of productivity improvement in low-permeability formations and unconventional reservoirs. It is done by injecting special fluids at pressure levels higher than formation breakdown pressure to develop artificial fractures that offer conductive pathways through which hydrocarbons can flow between the reservoir and the wellbore. Fracturing of modern conventional reservoirs uses advanced fluid systems and normally starts with pad stages of viscous, cross-linked polymer solutions to initiate and propagate fractures and then follows proppant-laden slurry stages to deliver and deposit particles of proppant in existing fractures[16,17].

The field application of guar-based polymer systems is mostly dominated by hydroxypropyl guar (HPG) cross-linked with borate or with metallic ions such as zirconium and titanium. The cross-linking process converts the linear polymer solutions into viscoelastic gels with significantly high viscosity and improved proppant suspension. The cross-linked borate systems exhibit fast viscosity formation and allow effective transport of the proppant, which poses a problem in predicting the friction pressure because of the rheology that varies with time[18]. Cross-linking kinetics are sensitive to pH, temperature, and polymer concentration, which adds variability to the operations.

Selection of proppant is a trade-off between economic and conductivity considerations. Sintered bauxite is a high-strength ceramic proppant suitable for high closure stress in deep, high-pressure formations but sells at premium prices. In the medium-stress levels, intermediate proppants that are prepared using kaolin offer decent service at lower cost. Sand is the most cost-effective, which is predominantly used in shallow formations or unconventional reservoirs where stresses at closure are not high[19]. The size of the proppant mesh is usually 16/30-100 mesh, and the size is determined mainly by the formation permeability and the required fracture conductivity[20].

2.2.1.2 Role of BHP in Fracture Propagation and Treatment Optimization

The evolution of fractures, such as fracture length, height, and width, is directly controlled by bottomhole pressure during fracturing[21]. Net pressure is the difference between the bottomhole pressure and minimum horizontal stress, which causes fractures to propagate and fracture width to be determined based on the principles of fracture mechanics in linear elasticity[22]. Low net pressure leads to small fracture sizes and ineffective placement of proppant, which reduces the effectiveness of treatment. High net pressure may cause an unregulated height increase to other formations, which decreases the efficiency of treatment due to the loss of fluid to the non-productive intervals[23].

Monitoring of bottomhole pressure in real time allows making critical decisions during the operation, such as changing the pumping rate, proppant concentration ramping schedule, and premature recognition of screenout conditions[24]. Screenout, i.e., proppant bridging in the fracture that inhibits further injection, is a major operational risk that needs instant attention in order to ensure that formation damage and equipment risks are minimized. The pressure profile before screenout is generally characterized by an accelerating bottomhole pressure rise, which, when detected early, gives time to terminate treatment before disastrous pump shutdown[25].

The bottomhole pressure data is subject to post-treatment analysis to give crucial information on the interpretation of fracture geometry. The analysis of pressure drop during shut-in stages after the treatment gives information about the dimensions of the fractures, the properties of fluid leakoff, and the efficiency of proppant placement[26]. The combination of pressure data and production performance allows validation of the fracture treatment models and optimization of the following treatments.

2.2.1.3 Challenges in BHP Prediction During Fracturing

There are inherent difficulties in the accurate prediction of the bottomhole pressure in the hydraulic fracturing process due to the complicated fluid rheology, proppant behavior, and dynamic operating conditions. Fracturing fluids applied in cross-linked modes are non-Newtonian and shear-thinning fluids that cannot be sufficiently described by simple power-law models. The kinetics of the rapid cross-linking of borate systems form time-dependent development of viscosity that changes with temperature exposure along the wellbore, making it difficult to compute friction pressure[27].

The mixture density and friction pressure are significantly elevated by the presence of proppant. Basic correction factors that are used in base fluid friction frequently do not suffice to characterize the complicated two-phase flow dynamics of proppant-laden slurries[28]. Non-homogeneous concentration profiles due to proppant settling, especially at lower flow velocities or using higher-density proppants, do not fit the assumptions of most of the correlation-based methods.

Fluid behavior is further complicated by the effects of temperature on fluid behavior. Geothermal gradients cause spatial variations in the wellbore thermal environment, whereas fluid injection cooling causes temporal variations in the wellbore thermal environment. The viscosity of cross-linked gel reduces with the increase in temperature, which influences the friction pressure and the capacity of carrying proppant. The instantaneous rheological state is dependent on the temperature history of fluid elements going through the wellbore, and thus coupled thermal-hydraulic modeling is necessary to predict the pressure accurately.

2.2.2 Gas Lift Well Systems

2.2.2.1 Principles of Gas Lift Technology

One of the most commonly used artificial lift techniques is gas lift, which is most effectively applied to high- and medium-gas-oil-ratio wells, wells with sand production issues barring downhole pumps, and in offshore service where surface space is scarce and the subsea installation is preferable[29]. The basic concept here is to inject compressed gas into the production tubing to decrease flowing bottomhole pressure to allow reservoir fluids to pass naturally to the surface against the lower hydrostatic head. Gas injection may be continuous or intermittent based on the well productivity as well as the availability of gas.

In continuous gas lift, the most widespread type, the gas injection continues at a constant rate by a subsurface injection valve at an optimum depth. The injected gas mixes the production fluid column, which decreases the mean mixture density and, as a result, lowers the hydrostatic pressure component. Several valve installations allow automatic optimization of the depth of gas injection, and the operating valve is based on the available gas injection pressure and wellbore pressure profile. Unloading valves that are above the primary operating valve help in starting up the well the first time by increasing the depth of injection as the liquid level drops[30].

The gas lift performance curve is the relationship between the rate of production and the rate of gas injection, which usually has an optimum injection rate that maximizes the rate of liquid. At low injection, there is not enough gas volume that is sufficient to lighten the column of fluid. Beyond the optimum injection, high friction losses due to the excessive velocity of the gas override the reduced density advantage. The shape and position of the performance curve depend on reservoir inflow performance, wellbore configuration, operating pressures, and fluid properties, including gas-oil ratio and water cut[31].

2.2.2.2 Importance of BHP in Gas Lift Optimization

Bottomhole pressure directly links gas lift system performance with reservoir productivity through inflow performance relationships. The reservoir deliverability, typically represented by productivity index, determines production rate as a function of pressure drawdown (reservoir pressure minus bottomhole flowing pressure). Gas lift optimization seeks to minimize bottomhole pressure while maintaining stable flow and respecting operational constraints, including available gas injection pressure and rate[32].

Knowledge of bottomhole pressure enables quantitative evaluation of gas lift efficiency[33]. The ratio of incremental liquid production to incremental gas injection, termed the gas lift efficiency, provides a metric for comparing operating points and identifying optimal injection strategies[34]. Operating at excessively high or low gas injection rates yields poor efficiency, unnecessarily consuming valuable lift gas that could enhance production from other wells sharing the same gas supply system[35].

Real-time bottomhole pressure monitoring facilitates detection of operational problems including gas lift valve malfunctions, formation damage progression, scale or asphaltene deposition, and liquid loading in deviated sections[36]. Unexpected increases in bottomhole pressure indicate reduced well productivity requiring investigation and potential remediation. Pressure fluctuations may reveal unstable flow conditions, including heading or severe slugging, requiring injection rate adjustment or valve configuration modification[37].

2.2.2.3 Challenges in BHP Prediction for Gas Lift Wells

Prediction of bottomhole pressure in gas lift wells is faced with the challenges of multiphase flow phenomena that cut across a variety of flow regimes. The gas-liquid mixture behavior changes from bubble flow at high liquid rates and low gas fraction and slug flow, where the gas and liquid slugs alternate, to annular flow, where the liquid forms a wall film with the gas flowing in the core. All the regimes have different characteristics of pressure gradients that need different forms of correlation or models[38].

The resulting flow patterns are highly sensitive to the operating conditions because of the strong interdependence between gas injection conditions and flow patterns. Variations in the rate of injection of gas cause variations in the gas-liquid ratio profiles along the wellbore, which may cause flow regime transitions that present discontinuities in the pressure gradient[39]. The effect of water cut on the properties of mixtures, such as density, viscosity, and surface tension, can have an effect on both the hydrostatic and friction pressure components[40]. The interactions between different gas-oil ratios and water cuts provide a high-dimensional space of parameters that empirical correlations find difficult to cover with adequate preciseness[41].

Change in temperature over the course of the wellbore has a major influence on the properties of the gas due to density and viscosity changes. Reductions in pressure cause gas expansion, which develops spatially varying velocities that affect losses to friction as well as the stability of flow regimes[42]. The non-isothermal condition of wellbore flow necessitates coupled thermal-hydraulic analysis in order to have the correct prediction of pressure, especially when dealing with deep wells or geothermal-heated conditions. This dynamic nature of the behavior after changing the rate of injection of the gas makes it difficult to perform real-time prediction of pressure because steady-state correlations cannot capture the dynamic effects[43].

2.2.3 Nitrogen Lifting Operations

2.2.3.1 Overview of Nitrogen Lifting Technology

Coiled tubing Nitrogen lift has been used as a temporary artificial lift for well cleanup, production testing, and reservoir evaluation operations. The gas injected, nitrogen, is chosen due to its inert properties and due to its comparative low cost when compared to hydrocarbon gases. This is injected at an optimal depth within the production tubing via coiled tubing. The injected nitrogen aerates the fluid column, which decreases the hydrostatic pressure, allowing the reservoir fluids to move to the surface[44]. This method is especially useful in the unloading of the completion fluids in new fractured wells, pulling out the liquid loading in a gas well, and in production tests without using permanent artificial lift equipment.

Coiled tubing deployment provides both flexibility in depth positioning and allows circulation operations in case they are necessary. The most common coiled tubing sizes are between 0.75 and 4.5 inches outside diameter, and the choice is dictated by the depth of the well, production tubing size, the fluid production rates expected, and the capacity of a nitrogen pump at hand. Coiled tubing outer diameter and production tubing inner diameter are separated by an annular space that is the main flow path of the reservoir fluids and injected nitrogen mixture. The geometry establishes unique flow properties as opposed to the traditional gas lift via specialized injection strings.

2.2.3.2 Importance of BHP in Nitrogen Lifting

The performance of the nitrogen lifting operations depends on bottomhole pressure at coiled tubing depth. A low-pressure drop does not cause reservoir inflow or meet set production rates. Too much pressure drawdown can cause damage to the formation by fines migration, proppant flowback of recently fractured wells, or flow instabilities such as severe heading that makes surface separation difficult[45]. The best bottomhole pressure is that which strikes a balance between the maximization of productivity and the maintenance of formation integrity and the limit of the capabilities of surface equipment.

Real-time bottomhole pressure monitoring permits a dynamic manipulation of the rate of nitrogen injection as well as the depth of coiled tubing to ensure that the optimum operating conditions are adhered to as the circumstances in the reservoir change. When conducting well unloading, the bottomhole pressure tends to decrease as the liquid level decreases and reservoir fluids start to flow into the wellbore. Monitoring this change of pressure helps in making decisions on changes in the rate of injection and when stable production has been achieved.

Bottomhole pressure plays a critical role in the economic optimization of nitrogen lifting operations[46]. The direct effect of the consumption of nitrogen is related to the cost of operation, and the average rates of nitrogen consumption will lead to costs of several thousands of dollars daily. To reduce the use of nitrogen and meet operational targets, there is a need to have a tight control of the bottomhole pressure by ensuring an optimal injection rate and coiled tubing depth[47]. The requirements of safety also demand that the bottomhole pressure be taken into consideration so that the uncontrolled proppant flowback is not caused by excessive drawdown.

2.2.3.3 Challenges in BHP Prediction During Nitrogen Lifting

The process of bottomhole pressure prediction during nitrogen lifting is facing significant difficulties related to transient multiphase flow in annular configuration. The flow route between the outer diameter of the coiled tube and the inner diameter of the production tube forms a hydraulic diameter and flow area that would be significantly different from that of conventional tubing, necessitating correlation corrections or special models[48]. Models often simplify the non-circumferential variations in flow characteristics caused by eccentricity when coiled tubing contacts the tubing wall.

The time-varying characteristics of nitrogen lifting processes generate time-varying pressure distributions that can no longer be well captured by steady-state correlations[49]. The level of liquid decreases steadily as production commences during well unloading, resulting in continuously changing boundary conditions. The flow rates are varying because the reservoir response to reduction in bottomhole pressure varies. The temperature curve changes with cold nitrogen injecting to cool the wellbore and the mixture being heated by the reservoir fluids.

The uncertainty in fluid properties increases difficulties in prediction. The compositions of reservoir fluids can be inadequately defined before production, especially in exploration wells or new reservoir developments[50]. Water cut, gas-oil ratio, and oil gravity impact on properties of the mixtures and, as a result, pressure gradients. The creation of emulsions between oil and produced water is capable of making significant modifications to effective viscosity, especially in wells with viscous crudes[51].

2.3 Traditional BHP Prediction Methods

2.3.1 Physics-Based Mechanistic Models

The most rigorous modeling of bottomhole pressure is the mechanistic multi-phase flow models, which are based on the first-principle governing equations of mass, momentum, and energy conservation[52,53]. These models explicitly consider phase separation, interfacial momentum transfer, and flow regime-dependent closure relationships. Multiphase flow in pipes has some basic conservation equations in the form of coupled partial differential equations that determine spatial and temporal changes in pressure, velocities of phases, and phase fractions[54].

Flow regime identification is an important part of mechanistic modeling. Regime boundaries are predicted in terms of dimensionless groups such as Froude number, Weber number, and gas-liquid velocity ratios. Closure relationships between the interfacial friction and liquid holdup, as well as the pressure gradient, vary with each regime (bubble, slug, churn, annular, stratified)[55].

Even though mechanistic models are theoretically rigorous, they have practical limitations. Large input data such as detailed fluid properties (densities, viscosities, and surface tensions at operating pressures and temperatures), pipe geometry, and flow rates might not be available easily in the field operations. Closure relationships of interfacial friction and phase distribution involve mathematical coefficients that were fitted to experimental data, which is a source of uncertainty[56]. The computational cost of addressing coupled different equations makes it difficult to implement in real time[57].

2.3.2 Empirical Correlations

Empirical correlations generated out of laboratory tests and field observations offer computationally effective substitutes to mechanistic models[58]. These correlations describe pressure gradients as algebraic functions of easily measured parameters such as flow rates, fluid properties, pipe diameter, and inclination angle[59]. The process of development is usually carried out in the form of dimensional analysis to establish the appropriate dimensionless groups, and then regression analysis of experimental data is conducted in order to obtain correlation coefficients and functional forms[60].

The Hagedorn-Brown correlation was the result of a large amount of experimental data on vertical flow, which predicts the liquid holdup as a result of dimensionless velocity number, liquid viscosity number, pipe diameter number, and liquid number. It is revealed that the correlation undergoes high performance in oil wells under some conditions but presents systematic deviations when used with a high gas liquid ratio[61,62]. The Beggs-Brill correlation has been extended to inclined flow by considering inclination effects in the form of correction factors that are determined through experimentation[63].

The Orkiszewski correlation is a compounding of existing correlations in which various formulations are used based on the flow regime. This strategy appreciates that general correlations across all regimes compromise the accuracy of the generality[64]. Although regime identification in theory is superior to single-correlation methods, it also creates discontinuities at the boundaries of transition. The inherent weakness of empirical correlations is low generalizability outside of the calibration range[65].

2.3.3 Hybrid Approaches

Hybrid methods attempt to leverage complementary strengths of mechanistic and empirical approaches while mitigating individual weaknesses. One common strategy employs mechanistic models for aspects amenable to first-principles analysis (hydrostatic pressure from mixture density calculations) combined with empirical correlations for phenomena difficult to model from fundamentals (friction factors and liquid holdup). This selective application exploits mechanistic model physical consistency while avoiding computational expense[66,67].

Another hybrid approach calibrates mechanistic model closure relationships against field data rather than relying solely on laboratory-derived coefficients. The adjusted models maintain physical structure while improving accuracy for specific applications. This data assimilation strategy has gained prominence with the increasing availability of permanently installed downhole instrumentation. Automated history matching algorithms systematically adjust model parameters to minimize discrepancies between predicted and measured pressures[68].

Despite improvements over purely empirical or mechanistic approaches, hybrid methods inherit limitations from constituent components. Calibrated mechanistic models require substantial field data for parameter adjustment, limiting applicability to new wells lacking measurement history. The calibration process may overfit available data, producing models that perform poorly under conditions deviating from calibration ranges.

2.3.4 Limitations and Accuracy Issues

Field implementation in hydraulic fracturing, gas lifting, and nitrogen lifting processes demonstrates that there is continually a problem of accuracy using the classic prediction techniques. The failure to effectively measure the rapidly cross-linking fracturing fluids' rheology causes a systematic underestimation of the pressure due to friction, especially at high rates of proppant. The standard correlations that were established to describe viscosity growth in a simple power-law or Newtonian fluid cannot capture the time-dependent behavior of viscosity growth observed in borate cross-linked systems. The resulting bottomhole pressure errors may be several hundred psi[69,70].

The friction pressure effects of proppant are poorly modeled in all forms of correlation. Basic multiplicative correction factors that are used to model base fluid friction do not sufficiently explain non-linear dependencies of concentration and flow regime changes related to proppant transport. Assuming homogeneous proppant distribution does not hold at settling where gradients of concentration are generated, especially during medium velocity of flow at high density of ceramic proppant[71,72].

In the comparison of correlation predictions with real bottomhole pressures in various well conditions, gas lift pressure prediction has significant scatter. Experiments investigating numerous correlations on the same wells often have deviations over 10-15 percent of measured pressure, which would represent errors of 200-300 psi in most practical work. These inaccuracies come about due to poor characterization of variations in liquid holdup in response to changes in the water cut and rate of gas injection[73].

There are other prediction challenges associated with nitrogen lifting operations due to the geometry of the annular flow and transient conditions. Conventional correlations derived in conventional tubing flow do not consider the effects of eccentricity of the coiled tubing and the changing pattern of the flow during the unloading of the well. The time dependence of liquid level drop and the constantly changing gas-liquid ratios are the obstacles to the applicability of steady-state correlations. A common error of prediction of up to 20% has been noted, especially during the high transition periods between unloading and stable production[74].

2.4 Machine Learning in Petroleum Engineering

2.4.1 Evolution of Data-Driven Approaches in Oil and Gas

Over the past decades, the petroleum industry is experiencing a rapid trend of incorporating data-based methodologies, as the capacity of data acquisition, computing power, and the complexity of algorithms have been growing exponentially[75–77]. The initial uses were on statistical analysis of production data, using decline curve analysis and simple regression models to predict well performance. The increase of permanently installed downhole sensors, advanced surface measuring devices, and computerized oilfield programs dramatically increased the amount of available data.

The adoption of machine learning as a fully fledged field in computer science in the 1990s and 2000s had a slow effect on petroleum engineering practice. The first application was focused on the pattern recognition problems such as log facies classification, seismic attribute analysis, and predicting equipment failures. These supervised learning exercises showed machine learning outperformed rules-based expert systems on problems with complex and non-linear relationships poorly modeled using conventional statistical tools[78,79].

The revolution in deep learning that started in the 2010s with the help of graphics processing unit computing and large labeled datasets created new opportunities in applications for petroleum. Convolutional neural networks were shown to be impressive in image analysis tasks such as automated fault interpretation of seismic data and lithology of wellbore images[80,81]. Although the application rate of machine learning in petroleum engineering has increased significantly, it remains lower than in other sectors since the culture is conservative and focused on established technologies, regulatory mandates to have explainable models, and lack of data science knowledge.

2.4.2 ML Applications Across Petroleum Engineering Domains

The application of machine learning has penetrated virtually every field of petroleum engineering, but the adoption and maturity differ widely. Supervised learning algorithms have been used in petrophysics to predict permeability, porosity, and fluid saturations of well logs, often better than traditional empirical transforms[82,83]. The neural networks, which are trained on the core analysis data, identify the complex relationships between the response of the logs and the petrophysical properties that the simple cross plots do not capture[84,85].

Another active field of application is drilling optimization. The models of machine learning forecast rate of penetration based on the parameters of drilling processes, allowing the weight on bit, rotary speed, and flow rate to be optimized in real-time. Kick detection algorithms use anomaly detection algorithms to detect early warning signs of formation fluid influx. Stuck pipe prediction models are used to evaluate the trends in drilling parameters in order to predict problematic situations before they occur[86,87].

The application of reservoir engineering includes reservoir characterization, production forecasting, and optimization of enhanced oil recovery. Machine learning is used to predict facies modeling and property distribution in geostatistical methods as well. Production decline analysis applies neural networks and ensemble techniques to predict well performance. Machine learning-based optimization algorithms have the advantage of high-dimensional parameter space exploration in history matching[88–90].

Applications of production engineering can be artificial lift optimization, well integrity monitoring, and optimization of surface facilities. The machine learning algorithms are used to forecast the best operating parameters of electric submersible pumps, gas lift systems, and sucker rod pumps using real-time well conditions. Predictive equipment failures can be used to schedule maintenance processes in advance, minimizing unplanned downtime[91,92].

2.4.3 ML Applications in Pressure Prediction

Specific machine learning uses in predicting pressure have been developed in various circumstances such as wellbore hydraulics, predicting the pressure in reservoirs, and modeling equipment performance[93,94]. Early prediction of steady-state flowing pressure in vertical wells was done, showing that neural networks could compete with or surpass the accuracy of empirical correlations in addition to allowing broader fluid property and flow conditions. These investigations showed possibility but were usually based on quite small data sets of restricted geographical scales.

New studies have widened the scope to more difficult prediction issues of pressure. In hydraulic fracturing treatments, neural networks have been used to predict wellhead pressure to monitor and optimize it in real-time. The models that are trained on the preliminary stages of treatments are found to be able to predict pressure development in later stages. Continuous learning, in which models are updated over time as new data are available, is potentially promising to capture the effect of time.

Machine learning pressure prediction in gas lift optimization studies is a progressively important addition to general optimization systems[95]. Reinforcement learning algorithms can be trained on optimum gas injection plans by interpreting reservoir-wellbore models or real field operations. These applications are important because of the ability to predict bottomhole pressure accurately.

The area of application that best aligns with this study, which is the forecasting of real-time bottomhole pressure during hydraulic fracturing, gas lift, and nitrogen lifting procedures, is relatively underreported in the literature.

2.4.4 Neural Networks and Ensemble Methods

Artificial neural networks represent the most extensively studied machine learning architecture for regression problems, including pressure prediction. The multilayer perceptron, comprising interconnected layers of artificial neurons applying weighted sums and nonlinear activation functions, demonstrates universal approximation capabilities for continuous functions. Training via backpropagation and gradient descent efficiently optimizes weights to minimize prediction errors on training data[96].

Deep learning, employing neural networks with multiple hidden layers, has revolutionized machine learning across computer vision and natural language processing. However, petroleum engineering applications typically employ relatively shallow networks (1-3 hidden layers) given modest dataset sizes compared to consumer applications. The risk of overfitting increases with model complexity, particularly problematic when training data span limited operational ranges[97].

Optimization algorithm selection significantly impacts neural network training efficiency and final performance. Traditional stochastic gradient descent with momentum provides baseline capabilities but requires careful learning rate tuning. Adaptive methods including Adam, RMSprop, and AdaGrad automatically adjust learning rates based on gradient statistics. For smaller datasets amenable to full-batch training, quasi-Newton methods like L-BFGS demonstrate faster convergence.

Ensemble methods combine predictions from multiple base models to achieve superior accuracy and robustness[98]. Random forests construct ensembles of decision trees trained on bootstrap samples with randomized feature selection, providing inherent resistance to overfitting. Gradient boosting builds ensembles sequentially, with each new model targeting residual errors from the existing ensemble. Modern implementations, including XGBoost, LightGBM, and CatBoost, incorporate regularization and optimized algorithms, achieving state-of-the-art performance[99].

2.4.5 Symbolic Regression and Interpretable AI

Symbolic regression is a radically new machine learning model that learns mathematical equations to characterize data relationships automatically[100]. Symbolic regression is used instead of trying to optimize parameters in fixed functional forms such as traditional regression and uses genetic programming to evolve pools of candidate equations in simulated natural selection. It is a method, based on searching a large space of potential mathematical formulas, of combining elementary functions to find equations that trade accuracy versus structural simplicity.

Symbolic regression has an interpretability advantage that solves a major critique of machine learning in petroleum engineering, which is the black-box quality of complicated models. Neural networks that have thousands of parameters are not transparent in terms of the way the inputs are converted into predictions. Engineers who are uneasy with opaque algorithms as a foundation to make critical decisions prefer physics-based models that give direct relationships. Symbolic regression fills this gap by generating human-readable equations, which can be checked by domain experts on physical aspects.

Recent symbolic regression implementations are done with advanced methods that improve efficiency as well as equation quality. Pareto front strategies explicitly trade accuracy against complexity and have populations of non-dominated solutions across the accuracy-interpretability spectrum. The dimensional analysis integration makes sure that derived equations are consistent in units and that they do not have expressions that do not have physical interpretation. The diversity increases with parallel island populations, which decrease premature convergence to local optima.

Although having benefits, symbolic regression has issues such as the computational cost of searching large expression space and finding equations that would need domain-specific functions beyond elementary operations. The technique is most effective in those problems that have a relatively simple underlying physics that can be formulated by relatively small equations. However, effective applications cut across various fields of science and prove to be useful when interpretability has high value[101–103].

2.5 Comparative Analysis: ML vs. Traditional Methods

2.5.1 Accuracy and Generalization

Comparative studies evaluating machine learning against traditional methods consistently demonstrate superior predictive accuracy for complex, nonlinear problems[104]. Neural networks and ensemble methods capture subtle patterns and interaction effects that empirical correlations miss[105]. The ability to learn optimal transformations of input features eliminates the need for manual feature engineering based on domain expertise. High-dimensional problems with numerous interacting parameters benefit particularly from machine learning's inherent capability to model complex functional dependencies.

Generalization performance, accuracy on unseen data, represents the ultimate metric distinguishing effective models from those merely memorizing training examples. Cross-validation procedures and blind testing on completely independent datasets provide rigorous assessment methodologies. Machine learning models properly trained with appropriate regularization demonstrate robust generalization across operational conditions spanning training data distributions. However, extrapolation beyond training ranges remains challenging[106].

The generalization advantage of machine learning becomes apparent when comparing performance across diverse well populations. A single neural network trained on data spanning multiple fields, formations, and operational practices often outperforms field-specific empirical correlations individually optimized for narrower conditions. This universality eliminates trial-and-error correlation selection procedures traditionally required when applying standard methods to new wells.

2.5.2 Real-Time Applicability and Computational Efficiency

Computational requirements are vital in real-time application viability, both in training and prediction. Machine learning systems have asymmetric computational characteristics; they require enormous amounts of resources (hours to days depending on dataset size and model complexity) to train but exhibit remarkably high prediction inference speeds (milliseconds). This property is beneficial in operational deployment in which models are trained offline with historical data and are used to make predictions fast within time-critical operations[107].

The effectiveness of trained machine learning models to execute computations can allow them to be deployed to edge computing devices such as wellsite controllers so that very real-time autonomous optimization is possible without being tied to cloud capabilities[108]. Contemporary processors can run these operations with a millisecond level of latency. Conventional approaches occupy a continuum between computationally trivial correlations, which need only algebraic computations, on the one hand, and computationally intensive mechanistic models, which need an iterative solution, on the other hand[109].

2.5.3 Adaptability and Cost-Effectiveness

Machine learning models have better adaptability to changing working conditions with simple retraining processes. With the drilling of new wells, the ways of operation evolve, or the equipment configurations are altered, and with merely the introduction of new data in training sets, new models are generated that adapt to changed circumstances. Unlike traditional correlations, this adaptive capability requires no redevelopment to extend predictions across wider calibration ranges[110,111].

When it comes to the use of machine learning versus conventional methods, the comparison of cost-effectiveness is in favor of the former in cases where bottomhole pressure monitoring is needed. Downhole gauge deployment removal saves $50,000-150,000 per well, which is transformational economics when operating on a large-scale basis. Conventional ways of calculating costs by using surface measurements offer similar cost savings but with limitations to accuracy. Machine learning is cost-effective and also more accurate[112].

The scalability benefit can be seen when used on a large population of wells. One trained model is used to apply to all the wells with similar characteristics when compared to traditional methods, which may need well-specific correlation selection and validation. The marginal cost of rollout to other wells is almost zero; just simply repeat the model with well-specific input parameters[113].

2.6 Knowledge Gaps and Research Opportunities

Although there has been considerable advancement in machine learning applications in solving problems in petroleum engineering, there remains a lot of knowledge gaps in terms of bottomhole pressure prediction in various operational areas. A majority of the published studies use individual applications or study small populations of wells without extensive comparative analysis under varying operating conditions, fluid systems, and well configurations. The comparative performance of various machine learning structures has not been sufficiently defined to be used in the application of pressure prediction. Strict comparison with conventional approaches on the same data under controlled conditions is mostly lacking.

The interpretability issue is one of the key impediments to the broader machine learning implementation in petroleum applications. Although symbolic regression has good opportunities in the direction of explainable models, the use of the method in predicting bottomhole pressure is still understudied. The trade-offs between interpretability and accuracy, i.e., a quantification of the amount of predictive performance that must be lost to obtain transparent mathematical expressions, must be systematically studied.

The methodologies of assessment of generalizability must be standardized and rigorous. The vast majority of research assesses performance with traditional train-test or k-fold cross-validation and does not give much information on what performance is like in entirely new settings. Stricter generalization tests are done with blind testing on entirely independent well populations of other fields. The reliance of the model performance on the size, diversity, and quality of the training dataset is a topic that should be investigated systematically.

Real-time implementation factors such as computational latency, software integration needs, and operations are yet to be examined properly in scholarly literature. The disconnection between tests in the laboratories on offline datasets and commercial applications in real field implementation is divided along technical, organizational, and cultural lines. Studies on the practical deployment issues, model updating processes, performance monitoring, failure mode identification, and human-machine interaction design would expedite the research-practice transfer of technology.

Chapter 3: Methodology

3.1 Research Framework and Approach

3.1.1 Overall Methodology Design

The present study uses a detailed data-driven framework to build machine learning models of bottomhole pressure prediction in three applications. The methodology will be systematic and consist of five stages (Figure 1): data collection, feature ranking, data preprocessing, model development and training, and strict validation. The structured approach provides consistency, reproducibility, and scientific rigor to all application domains but opens them to optimizations based on application-specific needs.

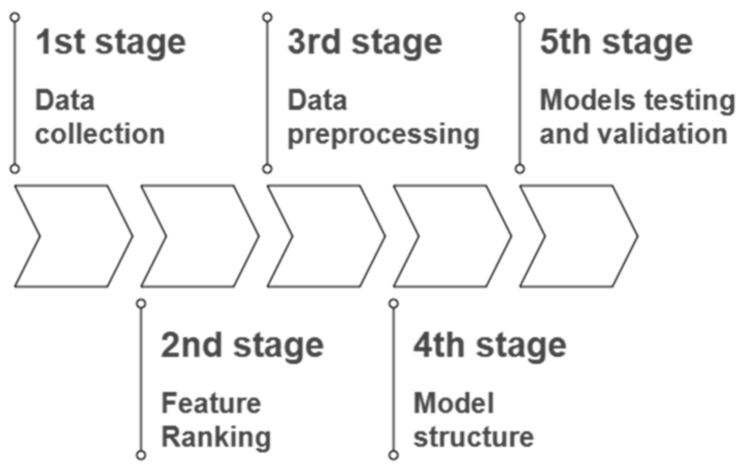

Figure 1. Schematic representation of the integrated machine learning methodology applied across hydraulic fracturing, gas lift, and nitrogen lifting applications. The framework comprises five sequential stages ensuring systematic development and rigorous validation of bottomhole pressure prediction models.

3.1.2 Integration of Three Application Domains

The study model incorporates the hydraulic fracturing, gas lift, and nitrogen lifting processes into a single machine learning pipeline. Although the operation parameter and physical aspect of each application are different, the common aspect is the difficulty of precise prediction of BHP. Such an integration indeed allows the comparative analysis of the performance of the models in various multiphase flow regimes, operation regimes, and well configurations. The framework takes advantage of interrelationships among applications and does not ignore the domain-specific requirements, which leads to the overall view of the ML-based pressure prediction capabilities within the real processes.

3.1.3 Workflow Overview

The workflow starts by collecting extensive field data in the vertical wells located in the western desert of Egypt that consists of 42 hydraulic fracturing wells, 304 gas lift wells, and 518 nitrogen-lifted wells. The correlation analysis (Pearson and Spearman) is done after data collection to determine important parameters of inputs to each application. Min-max normalization brings the features to the range of 0-1, which makes the model perform optimally. Various machine learning models, such as the linear models, the tree-based models, the ensemble model, the neural networks, and the symbolic regression, are developed and trained in a systematic way with the help of an 80/20 train-test split. MSE, RMSE, MAE, and R^2 are used to estimate model performance, and the validation based on k-fold cross-validation, repeated random sampling, and blind testing on independent datasets is used to verify the generalizability and applicability of the model in the field.

3.2 Data Collection and Field Context

3.2.1 Hydraulic Fracturing Dataset

The hydraulic fracturing data used in this project was assembled using 42 vertical wells in the western desert of Egypt. Borate cross-linked guar-based fracturing fluid, in particular, hydroxypropyl guar (HPG) polymer, was used to stimulate all of the wells. The data set includes an inclusive set of operational parameters, completion specifications, and treatment variables that are regularly measured and captured throughout the hydraulic fracturing operations.

The parameters measured are as follows: surface pressure (SP), slurry flow rate (SFR), surface proppant concentration (SPC), tubing inside diameter (TID), pressure gauge depth (PGD), gel load (GL), proppant size (PZ), and proppant specific gravity (PSG). The choice of these input variables was based on the basic principles of fluid mechanics that dictate the bottomhole pressure that is largely dominated by three key factors: the surface pressure, the hydrostatic pressure, and the frictional losses in pressure. The actual bottomhole pressure (BHP) recorded by downhole memory gauges used in the fracturing operations was used as the target variable in model development.

A statistical analysis of the dataset shows that there is a considerable difference in all the parameters, as shown in Table 1. The surface pressure was between 3591 and 9940 psi, with a mean of 6963 psi and a median of 7043 psi. The bottomhole pressure readings ranged between 3,842 and 15,774 psi, with an average of 10,490 psi and a median of 10,534 psi. Slurry flow rates were 22 to 37 barrels per minute, and surface proppant concentration was 0 to 7 pounds per gallon. The tubing internal diameter ranged between 2.992 and 3.826 inches, and depths of pressure gauges were between 6,132 and 13,140 feet, and the mean pressure gauge depth was 10,243 feet.

In this research, high-strength proppant (HSP) and intermediate-strength proppant (ISP) were considered: different mesh sizes, including 16/30 to 20/40, and specific gravities, including 3.265 to 3.547. The gel load was between 35 and 50 pounds per thousand gallons dependent on the design of treatment and formation needs.

Table 1. Statistical summary of the hydraulic fracturing dataset from 42 vertical wells showing minimum, maximum, average, and median values for all input parameters and target variable (bottomhole pressure).

Parameter	Units	MIN	MAX	AVG	Median
Surface pressure	psi	3591	9940	6963	7043
Bottomhole pressure	psi	3842	15,774	10,490	10,534
Slurry flow rate	bbl/min	22	37	30	30
Surface proppant concentration	lb/gal	0	7	3	3
Tubing inside diameter	inches	2.992	3.826	3.420	3.826
Pressure gauge depth	ft	6132	13,140	10,243	10,411
Gel load	lb/mgal	35	50	43	45
Proppant size	mesh	20/40	16/30	20/40	20/40
Proppant specific gravity	-	3.265	3.547	3.438	3.544

The visualization of the pair plot was done to analyze the pairwise relationships between the key variables (Figure 2), and it depicted that some parameters were significantly correlated. Surface pressure, bottomhole pressure, and pressure gauge depth in particular showed linear trends, which is expected to have physical dependencies. Categorical variables like proppant size and specific gravity were observed as discrete sparse points in the visualization space.

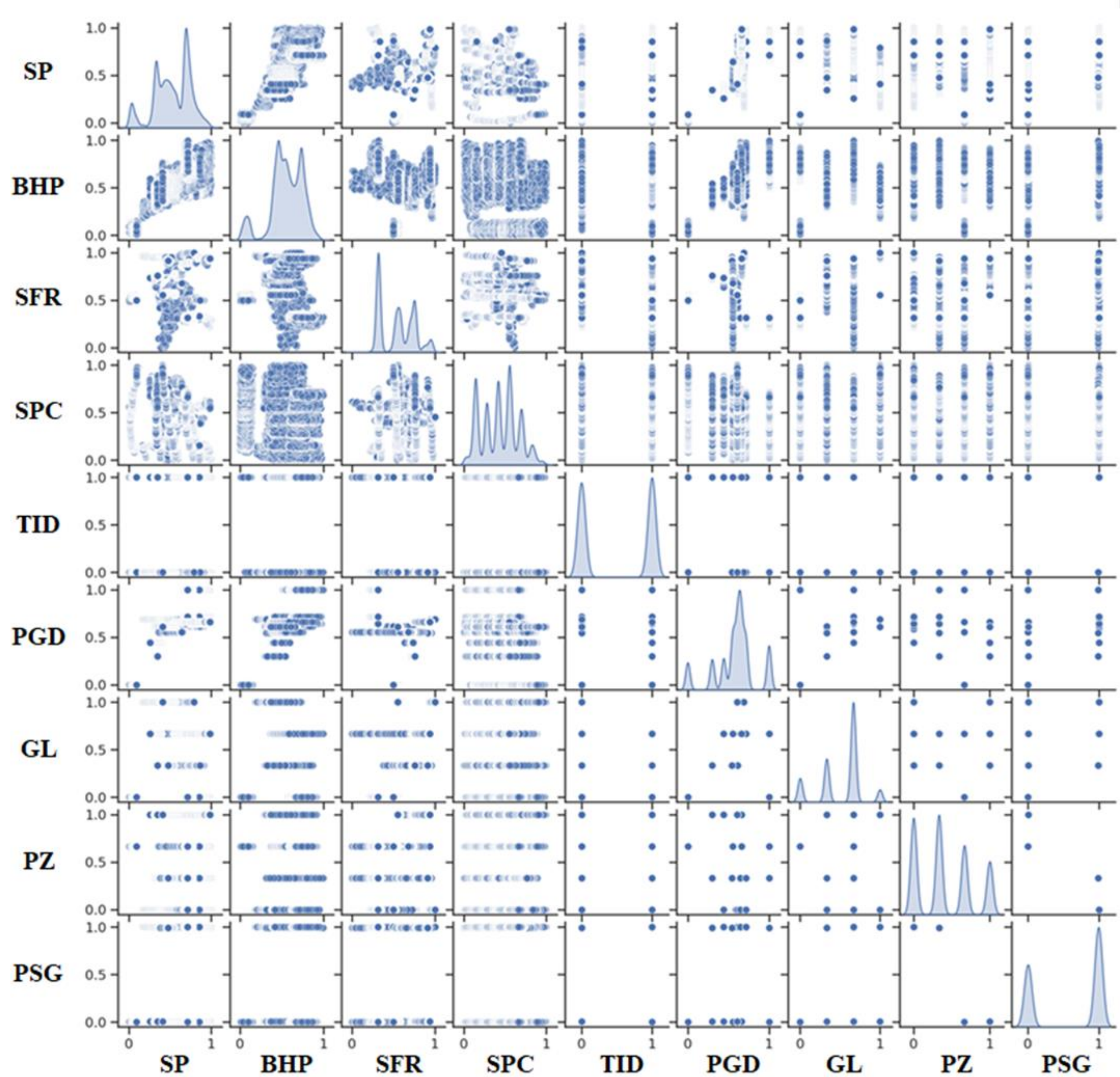

Figure 2. Pair plot matrix illustrating pairwise relationships and univariate distributions among key variables in the hydraulic fracturing dataset, with diagonal subplots showing individual parameter distributions and off-diagonal scatter plots revealing correlations, particularly notable linear trends between surface pressure, bottomhole pressure, and pressure gauge depth.

The violin plots used in the distribution analysis (Figure 3) gave further details regarding the variability of parameters and the pattern of concentration of data. Surface pressure and bottomhole pressure were distributed symmetrically and densely with very little variation around the mean stable values, indicating the similarity in operational practices across wells. The flow rate of the slurry was characterized by a narrow, peaked distribution that represented close-clustered values of operation. The tubing inner diameter showed bimodality with 2 different density peaks, which are two main completion configurations used in the field. The gel load parameter was more widely distributed with more variance, whereas proppant size and specific gravity had narrow distributions that were more concentrated with standardized proppant selection practices.

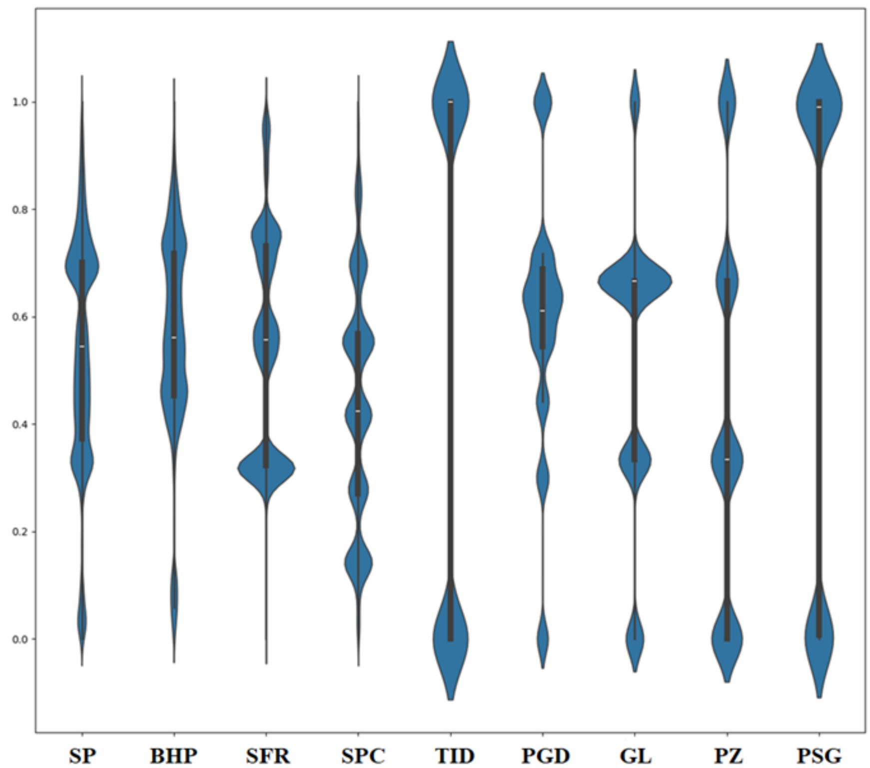

Figure 3. Violin plots displaying the distribution characteristics of normalized input parameters (scaled 0-1) for bottomhole pressure prediction in hydraulic fracturing operations, with plot widths indicating data density, white center lines representing medians, black bars denoting interquartile ranges, and individual points showing outliers, revealing symmetric distributions for pressure parameters, bimodal patterns for tubing inside diameter, and varying degrees of variability across operational variables.

This rich and holistic data allowed the creation of strong machine learning models that could describe complex, non-linear interactions among the parameters of operation and bottomhole pressure under different hydraulic fracturing conditions.

3.2.2 Gas Lift Wells Dataset

The gas lift wells data set is a compilation of 304 vertical wells that were operating in the western desert of Egypt. In this study, all wells were produced from multiple formations with a wide variety of operational conditions, which reflected a wide range of reservoir characteristics, completion specifications, and gas lift system configurations. This heterogeneity was critical in building strong regression-based machine learning models that have the ability to model complex relationships and make accurate predictions in various operational conditions.

The data set contains eleven major parameters measured regularly when performing gas lift operations. The input variables are tubing inside diameter (TID), tubing depth (TD), perforation depth (PD), wellhead pressure (WHP), reservoir temperature (RT), gas-oil ratio (GOR), water cut (WC), fluid flow rate (FFR), operating gas injection pressure (OGIP), gas lift injection rate (GLIR), and injection point depth (IPD). The model development target variable was the flowing bottomhole pressure (BHP) at perforation depth, which was measured using downhole pressure gauges installed through wireline on the routine gas lift flow surveys to determine the performance of the well under normal operating conditions.

The statistical analysis of the data obtained indicates that there is significant variation in data between parameters, as presented in Table 2. The inside diameter of the tubing was 2.441-2.992 inches, with a mean of 2.731 inches and a median of 2.992 inches. The depth of tubing was within the range of 3,024 to 10,494 feet with a median of 7,506 feet. The depth of perforation ranged between 3,052 and 10,544 feet. Wellhead pressure measurements were between 30 and 250 psi with large variability in operations with an average of 109 psi and a median of 100 psi.

The temperature of the reservoir was between 110.52 and 185.44, with an average temperature of 151.16. The gas-oil ratio ranged between 100 and 1,600 scf/stb with the mean of 654.96 scf/stb and median of 600 scf/stb. Water cut exhibited the largest scale, between 0 and 99 percent, with an average of 71.21 and a median of 81.5, representing wells at varying levels of their productive life span. Gas injection pressure was between 800 and 1,100 psi, with gas lift injection rates between 0.3 and 1.2 mmscf/d. The bottomhole pressure readings were spread between 415 and 1,669 psi, and the mean was 830 psi. The depth of injection points was between 2,304 and 9,390 feet, and the fluid flow rates ranged between 22 and 1,963 stb/d.

Table 2. Statistical summary of the gas lift wells dataset from 304 vertical wells showing minimum, maximum, average, and median values for all operational parameters, completion specifications, and target variable (bottomhole pressure at perforation depth).

Parameter	Units	MIN	MAX	AVG	Median
Tubing inside diameter	inches	2.441	2.992	2.731	2.992
Tubing depth	ft	3002	10,494	7022.441	7506
Perforation depth	ft	3052	10,544	7116.322	7656
Wellhead pressure	psi	30	250	109.407	100
Reservoir temperature	f	110.52	185.44	151.163	156.56
Gas–oil ratio	scf/stb	100	1600	654.960	600
Water cut	%	0	99	71.2105	81.5
Operating gas injection pressure	psi	800	1100	912.5	900
Gas lift injection rate	mmscf/d	0.3	1.2	0.566	0.5
Bottomhole pressure	psi	415	1669	830.286	803.5
Injection point depth	ft	2304	9390	6099.862	6379.5
Fluid flow rate	stb/d	22	1963	730.256	676

The pair plot graph (Figure 4) was used to demonstrate the relationships between key variables in the data set. Significant positive linear relationships were found between bottomhole pressure and a number of variables such as tubing depth, perforation depth, reservoir temperature, water cut, and fluid flow rate, indicating anticipated physical dependencies. There were also good linear relationships between tubing depth, reservoir temperature, injection point depth, and perforation depth as expected in normal gas lift operations.

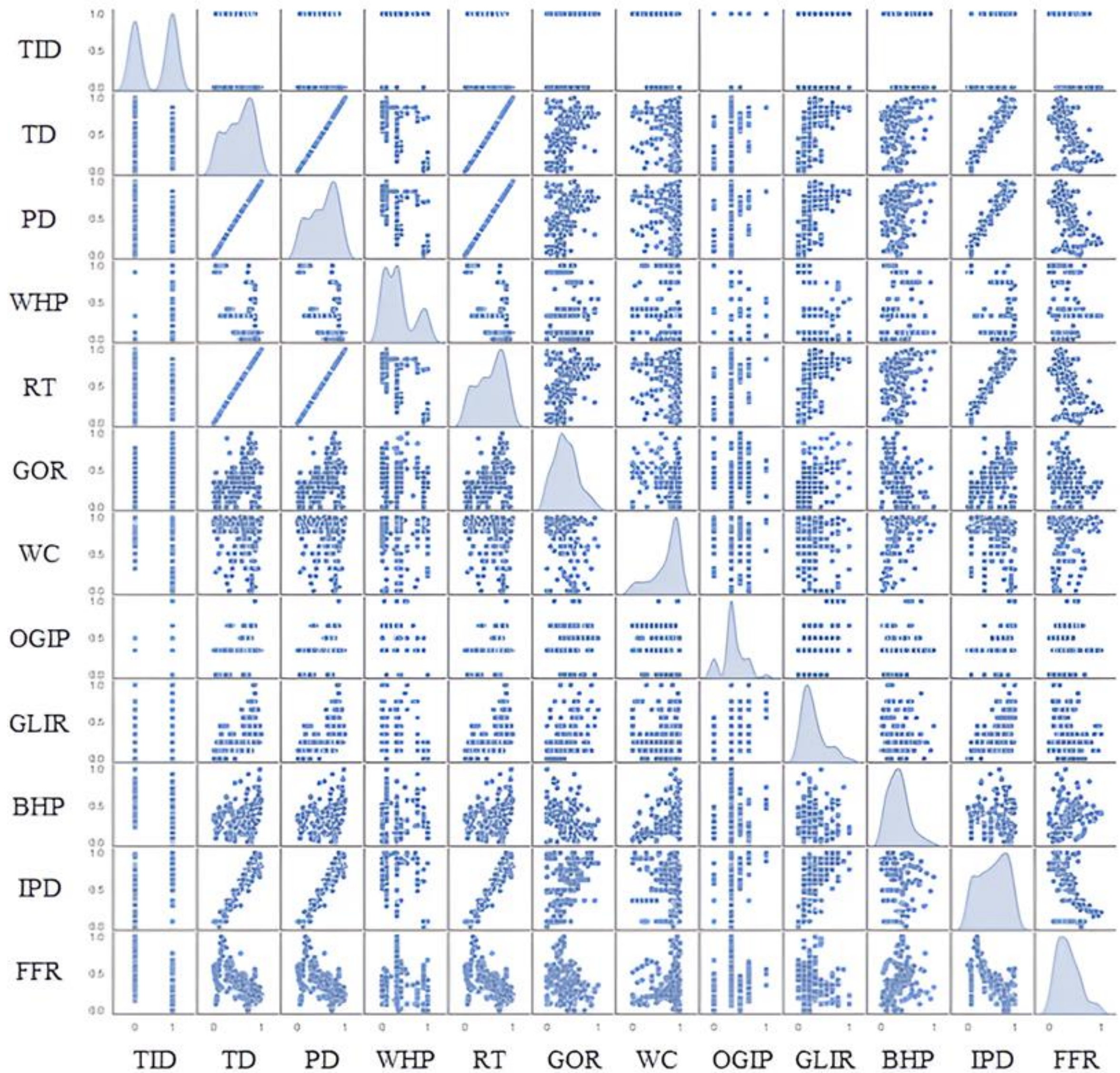

Figure 4. Pair plot matrix illustrating pairwise relationships and univariate distributions among all variables in the gas lift wells dataset, with diagonal subplots displaying individual parameter distributions and off-diagonal scatter plots revealing correlations, notably strong linear relationships among tubing depth, perforation depth, reservoir temperature, water cut, fluid flow rate and bottomhole pressure.

The use of violin plots to conduct distribution analysis (Figure 5) showed that the parameters had different patterns. The tubing's inside diameter had bimodal features, showing that there were two common completion configurations. Tubing depth, perforation depth, and reservoir temperature showed dense symmetric distributions with little variation. Bigger ranges with respect to operational flexibility were presented in wellhead pressure and gas lift injection rate. The pattern of gas-oil ratio and water cut showed closely clustered patterns, whereas operating gas injection pressure had a narrow peak. The bottomhole pressure was moderately, symmetrically distributed, and fluid flow rate was broadly distributed, indicating a high level of variability in the rate of production across wells.

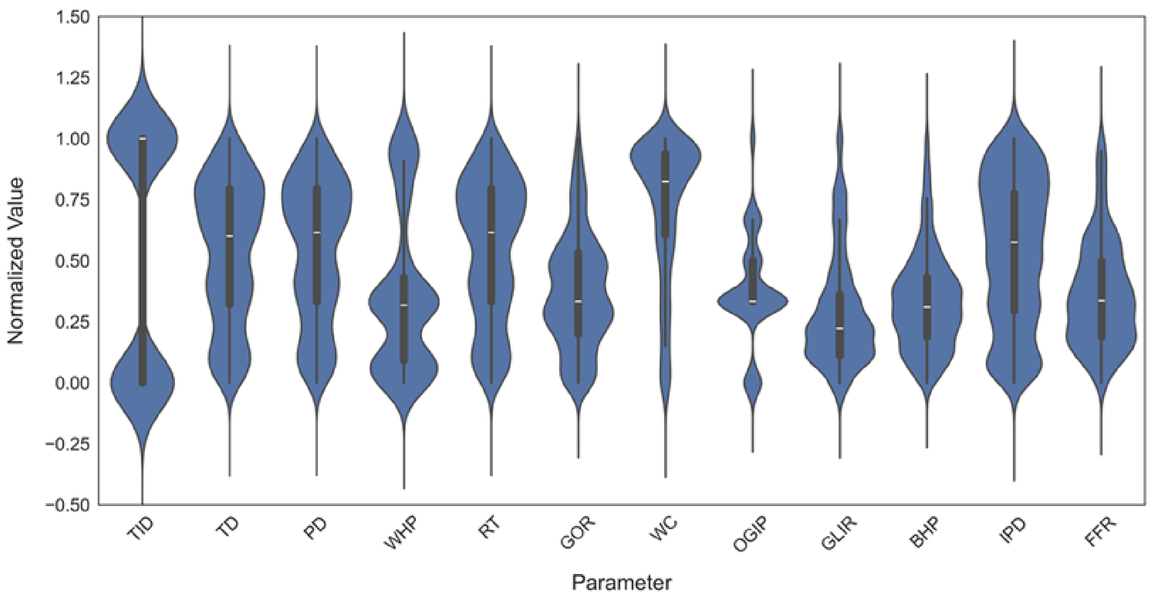

Figure 5. Violin plots displaying the distribution characteristics of normalized parameters (scaled 0-1) in the gas lift wells dataset, with plot widths indicating data density, white center lines representing medians, black bars denoting interquartile ranges, and individual points showing outliers, revealing bimodal distribution for tubing inside diameter, symmetric patterns for depth and temperature parameters, and varying degrees of operational variability across gas lift system variables.

3.2.3 Nitrogen-Lifted Wells Dataset

The nitrogen-lifted wells data set was compiled on 518 vertical oil wells in the western desert in Egypt. In all wells, coiled tubing nitrogen lifting operations were used to start or restart fluid production, especially where the wellbore was filled with completion fluids or formation fluids that inhibited natural production. The standardized tubing was used in completing the wells: 3.5-inch outer diameter, 2.99-inch inner diameter, and 9.3 lb/ft N-80 grade production tubing with External Upset Ends (EUE) thread, compliant with API 5CT standards. The coiled tubing operations used 1.5-inch outer diameter, 0.134-inch wall thickness, and 15,000-foot-long coiled tubing.

The data set includes ten operation and reservoir parameters that are regularly monitored in the process of nitrogen lifting operations. The input variables are fluid flow rate at surface (FFR-S), water cut (WC), gas-oil ratio (GOR), water salinity (WS), wellhead flowing pressure (WHP), wellhead flowing temperature (WHT), coiled tubing depth (CTD), nitrogen rate (NR), and oil gravity (OG). The dependent variable was the bottomhole pressure at coiled tubing depth (BHP-CTD), which was determined by downhole pressure gauges installed through coiled tubing during the normal nitrogen lift operation and production tests.

The statistical analysis showed that the parameters vary significantly, as in Table 3. The pressure at coiled tubing depth was between 158 and 5,942 psi with a mean of 2,141 psi and a median of 1,737 psi. The fluid flow rates ranged between 80 and 4,510 stb/d with an average of 1,552 stb/d. The water cut covered the full extent of 0-100 percent with a mean of 41 percent and a median of 50 percent. The gas-oil ratio was varied between 0 and 2,000 scf/stb, with an average of 609 scf/stb. The salinity of water was between 49,995 and 200,000 ppm. Wellhead flowing pressure was 13 to

570 psi, and wellhead flowing temperature was 72 to 160°F. The depth of coiled tubing was 3,000 - 13,040, with an average of 8,250. Nitrogen injection rates were 400 to 1,000 scf/m, and oil gravity was 12 to 54 API.

Table 3. Statistical summary of the nitrogen-lifted wells dataset from 518 vertical oil wells showing minimum, maximum, average, and median values for operational parameters, fluid properties, and target variable (bottomhole pressure at coiled tubing depth).

Parameter	Units	MIN	MAX	AVG	Median
Bottomhole pressure at coiled tubing depth	psi	158	5942	2141	1737
Fluid flow rate at surface	stb/d	80	4510	1552	1210
Water cut	%	0	100	41	50
Gas–oil ratio	scf/stb	0	2000	609	319
Water salinity	ppm	49,995	200,000	150,941	150,000
Wellhead flowing pressure	psi	13	570	80	62
Wellhead flowing temperature	f	72	160	103	102
Coiled tubing depth	ft	3000	13,040	8250	8971
Nitrogen rate	scf/m	400	1000	519	500
Oil gravity	API	12	54	37	35

The pair plot (Figure 6) showed that there were significant linear relationships between BHP-CTD and a number of parameters, most notably fluid flow rate at surface, water cut, and coiled tubing depth, which are expected to be physically dependent. Violin plots used in distribution analysis (Figure 7) indicated that the water salinity had a symmetric distribution and the concentration was clearly defined around the higher normalized values. Wellhead pressure and nitrogen rate exhibited small variations and small dispersion, whereas fluid flow rate and coiled tubing depth showed larger dispersion, suggesting that there is large inter-well variation.

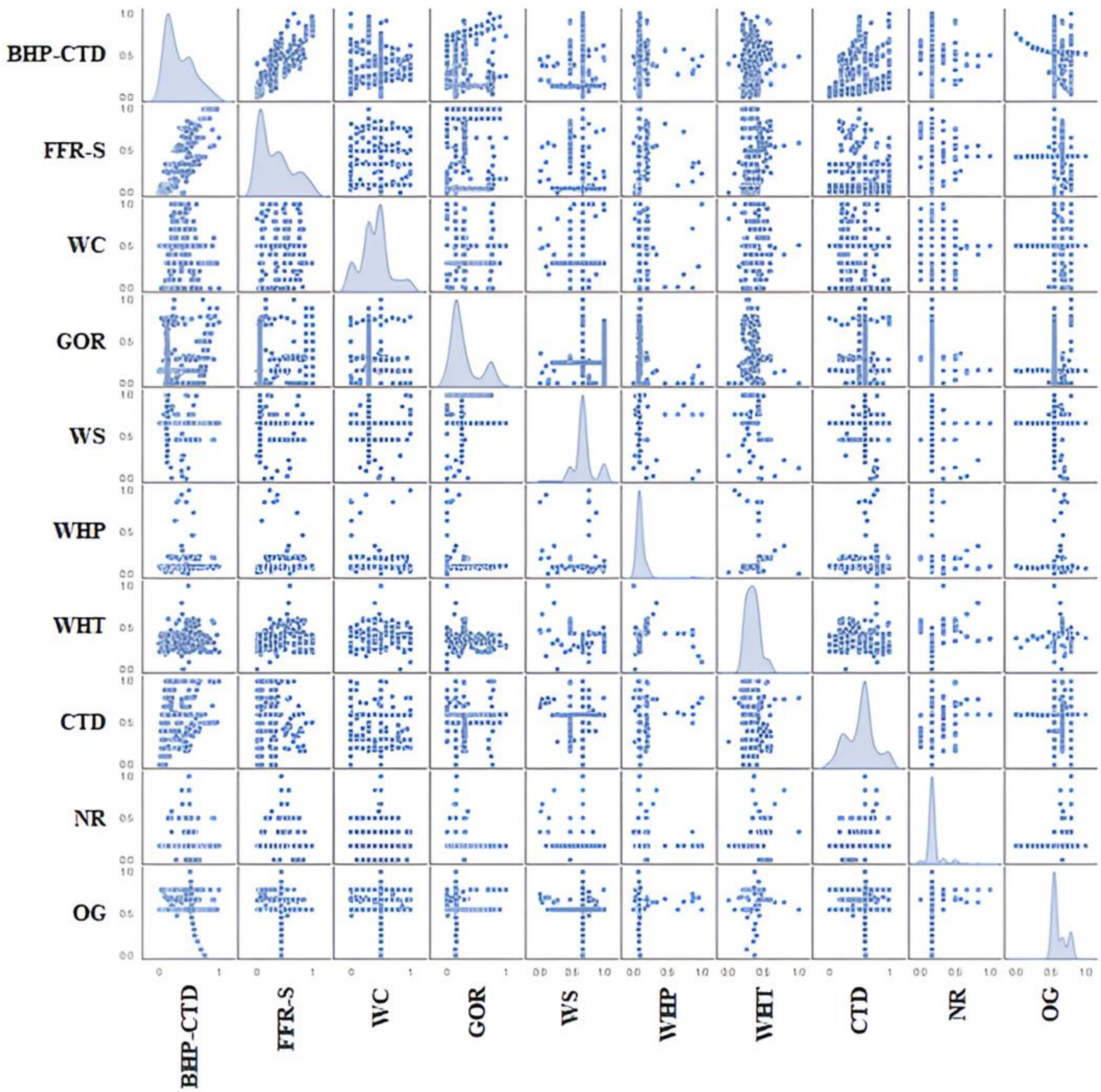

Figure 6. Pair plot matrix illustrating pairwise relationships and univariate distributions among operational parameters in the nitrogen-lifted wells dataset, with diagonal subplots displaying individual parameter distributions and off-diagonal scatter plots revealing correlations, particularly linear relationships between bottomhole pressure at coiled tubing depth and fluid flow rate at surface, water cut, and coiled tubing depth.

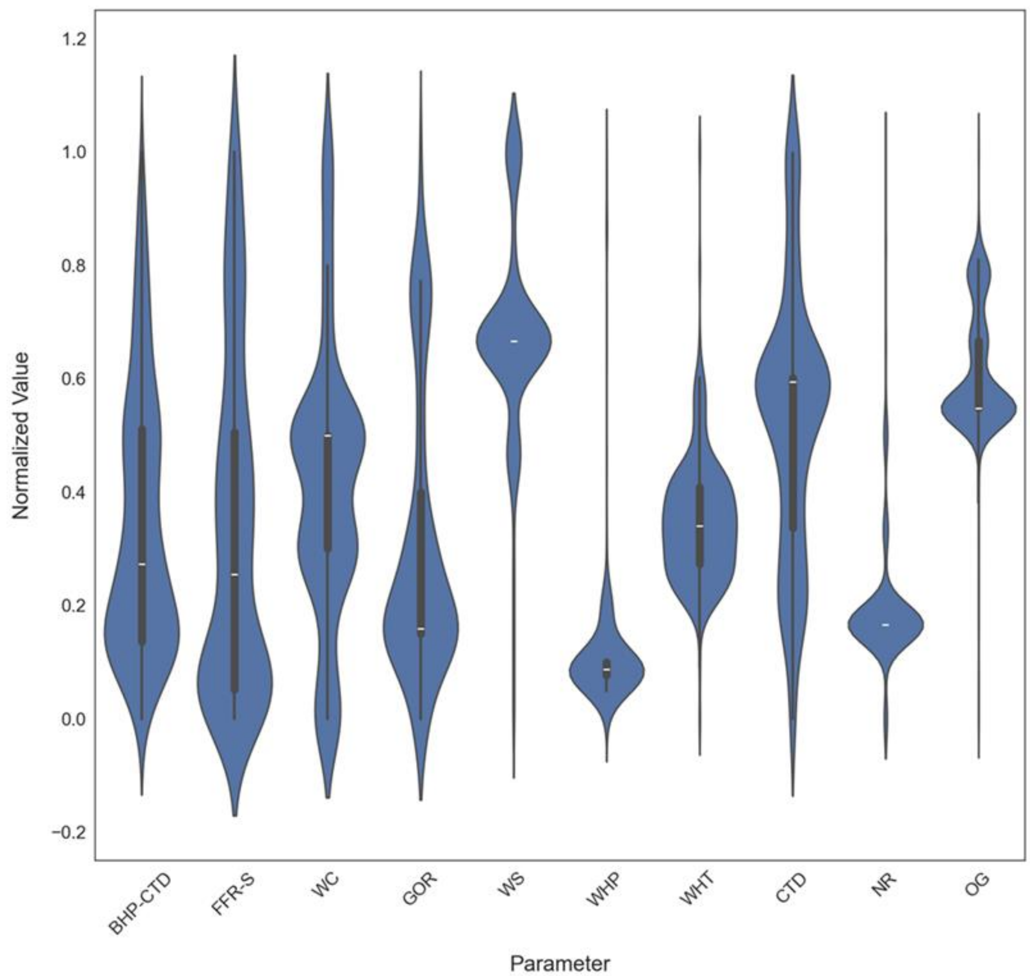

Figure 7. Violin plots displaying the distribution characteristics of normalized parameters (scaled 0-1) in the nitrogen-lifted wells dataset, with plot widths indicating data density, white center lines representing medians, black bars denoting interquartile ranges, and individual points showing outliers, revealing symmetric distribution for water salinity, narrow peaked distributions for wellhead pressure and nitrogen rate, multimodal patterns for water cut and gas-oil ratio, and broad distributions for fluid flow rate and coiled tubing depth reflecting operational variability.

3.3 Feature Ranking

3.3.1 Ranking of Hydraulic Fracturing Features

Understanding the relationships between input parameters and the target variable is fundamental to developing effective machine learning models for bottomhole pressure prediction. In the hydraulic fracturing application, feature ranking was conducted to identify which operational and completion parameters exert the most significant influence on BHP during fracturing operations. This analysis employed two complementary statistical methodologies: Pearson's correlation coefficient and Spearman's rank correlation coefficient, each offering distinct insights into the data structure.

Pearson's correlation coefficient (r) measures the linear relationship between two continuous variables, as expressed in Equation (4). This metric is particularly effective for identifying proportional relationships but demonstrates sensitivity to outliers and assumes normally distributed data. In contrast, Spearman's rank correlation coefficient (ρ), formulated in Equation (5), evaluates monotonic relationships using ranked data rather than raw values. This approach provides greater robustness against outliers and non-normal distributions, making it suitable for ordinal data or relationships that may not be strictly linear. Both coefficients range from −1 to +1, where +1 indicates perfect positive correlation, −1 represents perfect negative correlation, and 0 signifies no correlation.

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2} \cdot \sqrt{\sum(Y_i - \overline{Y})^2}} \tag{4}$$

where:

- X_i and Y_i are the individual data points;
- $\overline{X}$ and $\overline{Y}$ are the means of X and Y;
- The numerator is the covariance of X and Y;
- The denominator is the product of their standard deviations.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{5}$$

where:

- d_i is the difference between the ranks of corresponding values in X and Y;
- n is the number of data points.

The analysis of the feature ranking showed that the influence of the input parameters had different patterns. Figure 8 shows the relative significance of each feature according to the two methodologies of the correlation. Surface pressure (SP), pressure gauge depth (PGD), and slurry flow rate (SFR) became the parameters that had the most significant correlations with BHP, and the rest of the variables showed moderate correlations. This order of influence is an indication of the physics underlying the pressure interactions in hydraulic fracturing processes, where surface pressure is a direct factor in the downhole environment, gauge depth sets elements of hydrostatic pressure, and flow rate impacts frictional pressure losses.

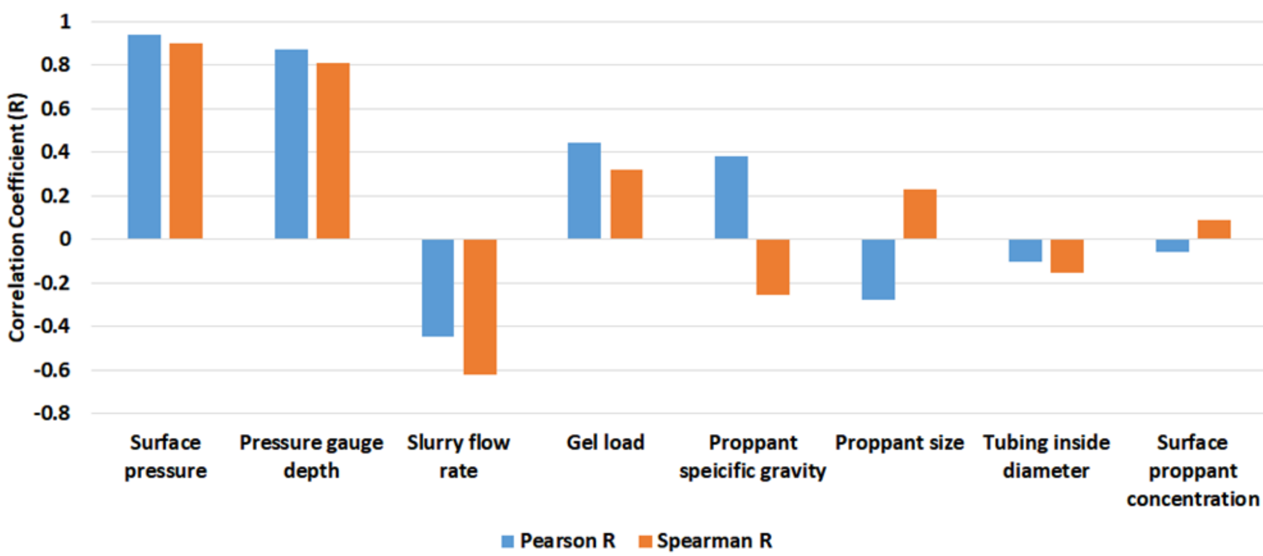

Figure 8. Feature importance ranking for hydraulic fracturing input parameters based on Pearson's and Spearman's correlation coefficients with bottomhole pressure. Surface pressure (SP), pressure gauge depth (PGD), and slurry flow rate (SFR) exhibit strong correlations with BHP, while other operational parameters including surface proppant concentration (SPC), tubing inside diameter (TID), gel load (GL), proppant size (PZ), and proppant specific gravity (PSG) demonstrate moderate correlations.

The Pearson correlation matrix, visualized in Figure 9, quantitatively demonstrates these relationships across all input parameters. Notably, surface pressure exhibited a near-perfect positive correlation with BHP (r = 0.94), while pressure gauge depth showed a strong positive correlation (r = 0.87). These findings confirm the dominant role of these parameters in determining bottomhole pressure conditions. Conversely, certain parameters such as tubing inside diameter and pressure gauge depth displayed weak correlations with each other, indicating relative independence and minimal multicollinearity concerns.

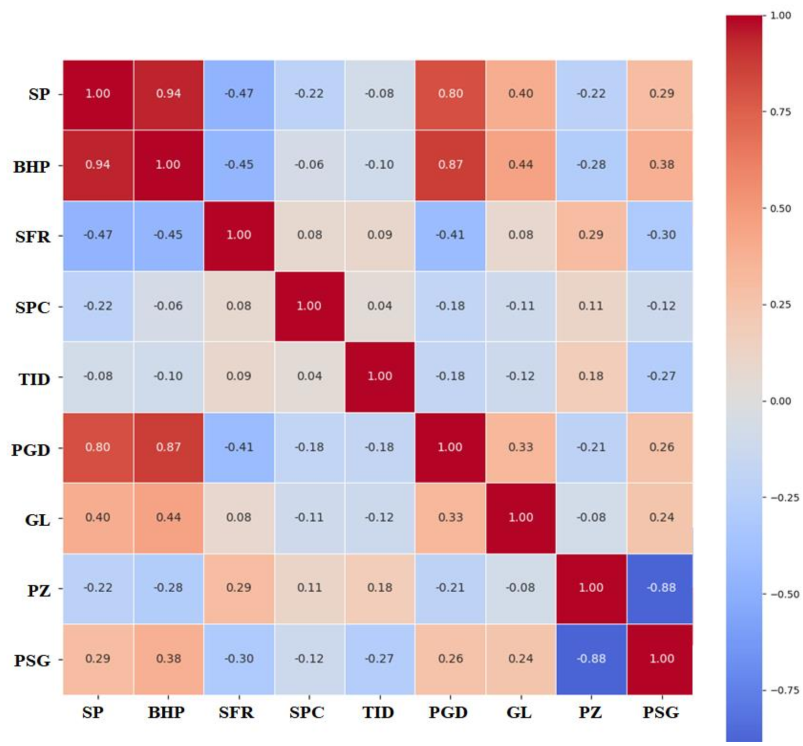

Figure 9. Pearson correlation coefficient heatmap illustrating linear relationships among input parameters in the hydraulic fracturing dataset. Surface pressure demonstrates near-perfect positive correlation with bottomhole pressure (r = 0.94), and pressure gauge depth shows strong positive correlation (r = 0.87), indicating strong interdependence.

Complementary insights emerged from the Spearman correlation analysis, presented in Figure 10. Surface pressure and pressure gauge depth maintained strong positive correlations with BHP (ρ = 0.90 and ρ = 0.81, respectively), while slurry flow rate demonstrated a moderate inverse relationship (ρ = −0.62). Other parameters, including surface proppant concentration, tubing inside diameter, gel load, proppant size, and proppant specific gravity, exhibited weaker correlations, suggesting more subtle or indirect influences on BHP.

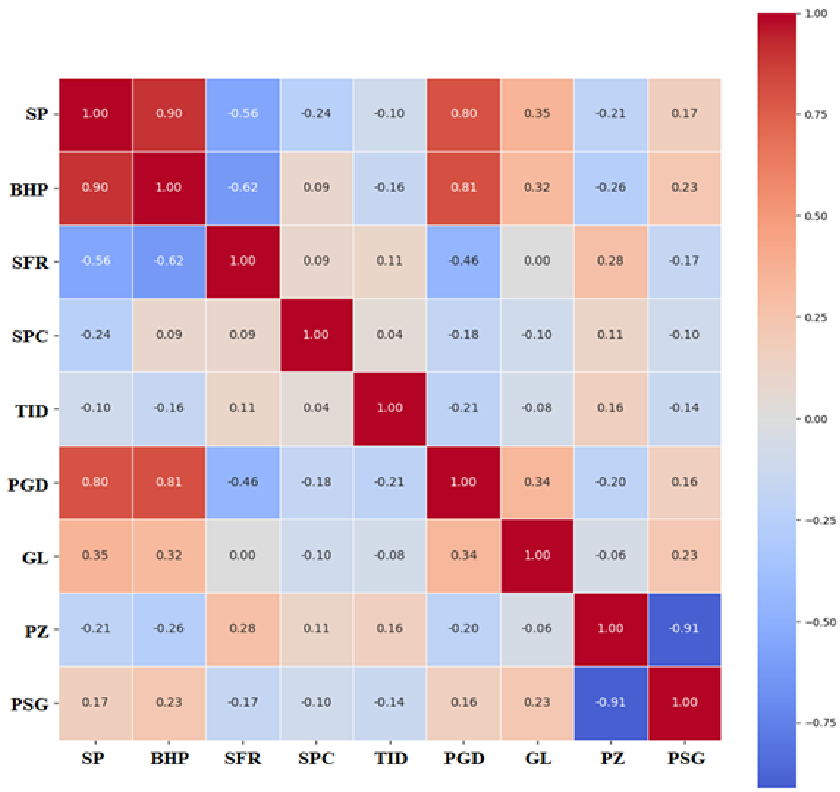

Figure 10. Spearman rank correlation coefficient heatmap displaying monotonic relationships between features in the hydraulic fracturing dataset. Surface pressure and pressure gauge depth maintain strong positive correlations with bottomhole pressure (ρ = 0.90 and ρ = 0.81, respectively), while slurry flow rate demonstrates moderate inverse relationship (ρ = −0.62). Other parameters show weaker correlations, indicating minimally direct influence on BHP prediction.

These correlation analyses offered valuable advice to future model development, as they confirmed the key physical factors of BHP are represented by the chosen input parameters and the fact that the input parameters are of manageable dimensionality and computationally efficient.

3.3.2 Ranking of Gas Lift Well Features

The same statistical correlation methodologies were used to rank the features to be used in the gas lift wells application to calculate the most significant parameters that influence BHP at the perforation depth. The operational variables covered in the analysis were tubing inside diameter (TID), tubing depth (TD), perforation depth (PD), wellhead pressure (WHP), reservoir temperature (RT), gas-oil ratio (GOR), water cut (WC), fluid flow rate (FFR), operating gas injection pressure (OGIP), gas lift injection rate (GLIR), and injection point depth (IPD).The ranking of feature importance presented in Figure 11 showed that the variables of water cut, tubing inside diameter, and fluid flow rate showed high correlation with BHP, and other variables showed moderate correlations. The Pearson correlation heatmap (Figure 12) measured these relationships, which showed BHP had moderate positive correlation with WC (0.52) and weaker with FFR (0.28), and negative with TID (−0.37). WHP, OGIP, and GLIR parameters demonstrated the least correlations with other variables, and this can be seen as a sign of relative independence of the parameters in the dataset.

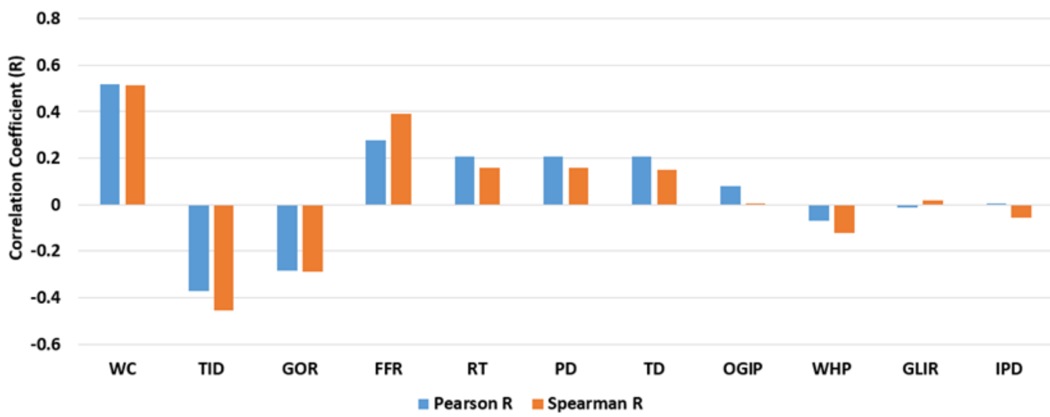

Figure 11. Feature importance ranking for gas lift well parameters based on Pearson's and Spearman's correlation coefficients with bottomhole pressure. Water cut (WC), tubing inside diameter (TID), and fluid flow rate (FFR) demonstrate the strongest correlations with BHP, while other operational parameters exhibit moderate influence.

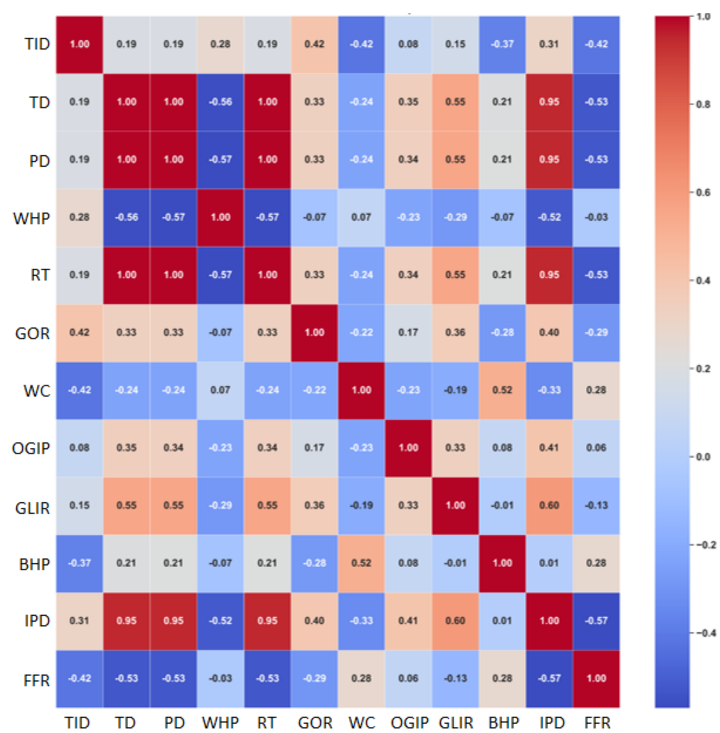

Figure 12. Pearson correlation coefficient heatmap illustrating linear relationships among input and output parameters in the gas lift wells dataset. Bottomhole pressure shows moderate

correlation with water cut (r = 0.52) and negative correlation with tubing inside diameter (r = −0.37).

The Spearman correlation analysis (Figure 13) reinforced these findings, with WC and FFR showing moderate positive correlations with BHP (ρ = 0.51 and ρ = 0.39, respectively), and TID displaying moderate negative correlation (ρ = −0.46). These correlation patterns provided essential guidance for understanding the complex multiphase flow dynamics governing BHP in gas lift operations.

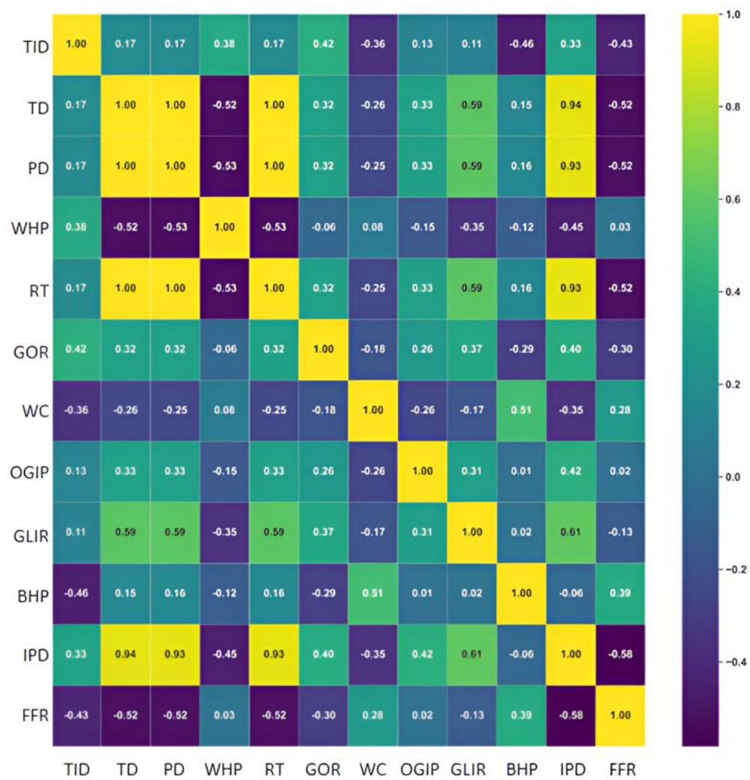

Figure 13. Spearman rank correlation coefficient heatmap displaying monotonic relationships between features in the gas lift wells dataset. Water cut and fluid flow rate exhibit moderate positive correlations with bottomhole pressure (ρ = 0.5 and ρ = 0.4, respectively), while tubing inside diameter shows moderate negative correlation (ρ = −0.5), indicating inverse relationship with BHP at perforation depth.

3.3.3 Ranking of Nitrogen-Lifted Wells Features

Feature ranking for nitrogen-lifted wells employed identical correlation methodologies to identify parameters most influential in predicting bottomhole pressure at coiled tubing depth (BHP-CTD). The analysis encompassed operational variables including fluid flow rate at the surface (FFR-S), water cut (WC), gas–oil ratio (GOR), water salinity (WS), wellhead flowing pressure (WHP), wellhead flowing temperature (WHT), coiled tubing depth (CTD), nitrogen rate (NR), and oil gravity (OG).

Figure 14 demonstrates the relative influence of each input parameter on BHP-CTD prediction. Fluid flow rate at the surface exhibited the strongest positive correlation with BHP-CTD, while remaining parameters demonstrated moderate correlations. The Pearson correlation heatmap (Figure 15) quantified these relationships, revealing a particularly strong positive correlation between BHP-CTD and FFR-S (r = 0.84). Other correlations proved substantially weaker or negligible, such as WC and WHP (r = −0.05), indicating minimal linear relationships.

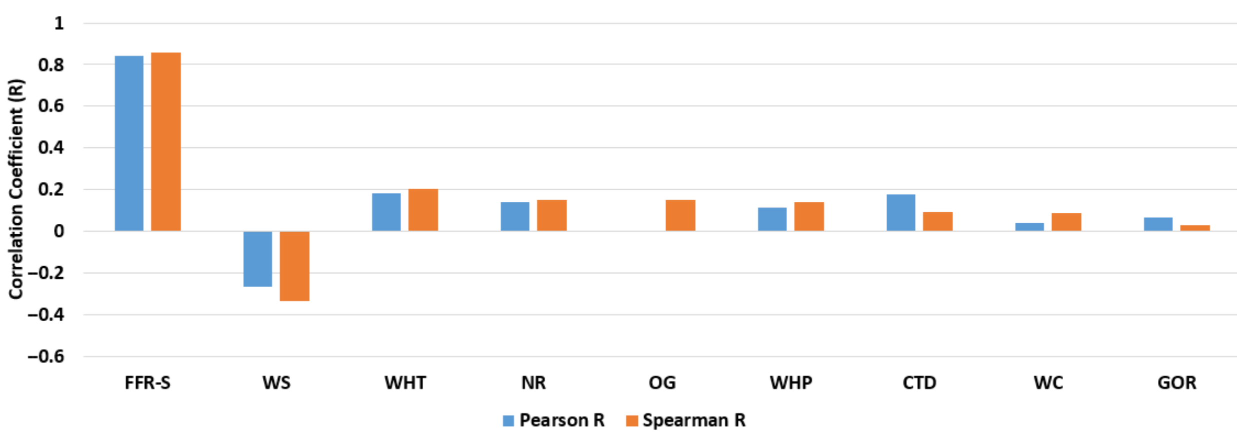

Figure 14. Feature importance ranking for nitrogen-lifted well parameters based on Pearson's and Spearman's correlation coefficients with bottomhole pressure at coiled tubing depth (BHP-CTD). Fluid flow rate at surface (FFR-S) demonstrates the strongest correlation with BHP-CTD, while other operational parameters including water cut (WC), gas–oil ratio (GOR), water salinity (WS),

wellhead flowing pressure (WHP), wellhead flowing temperature (WHT), coiled tubing depth (CTD), nitrogen rate (NR), and oil gravity (OG) exhibit moderate correlations.

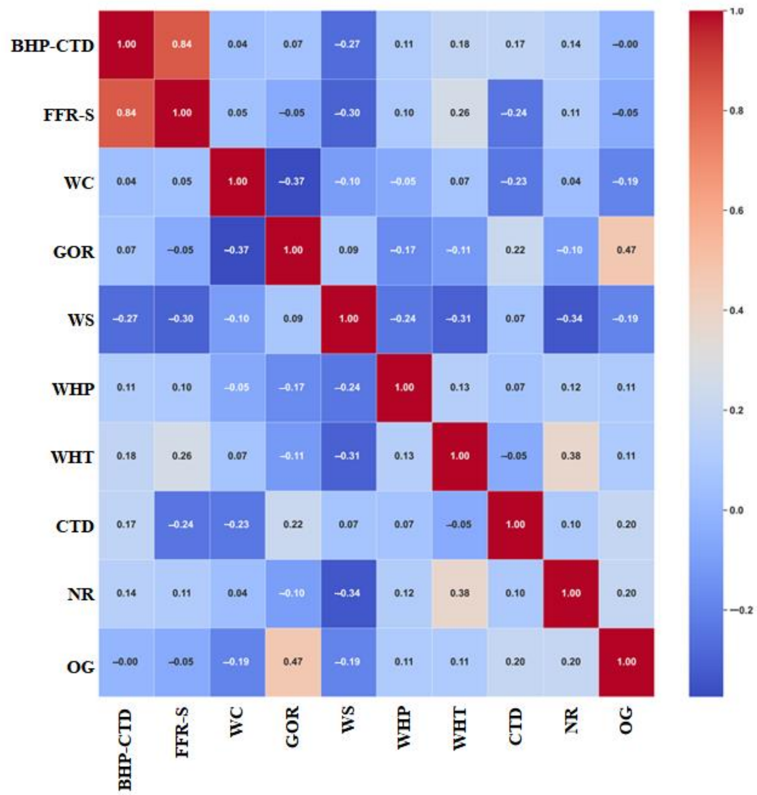

Figure 15. Pearson correlation coefficient heatmap illustrating linear relationships among operational parameters in the nitrogen-lifted wells dataset. Fluid flow rate at surface exhibits strong positive correlation with bottomhole pressure at coiled tubing depth (r = 0.84), while other parameter pairs demonstrate substantially weaker or negligible correlations, such as water cut and wellhead flowing pressure (r = −0.05), indicating relative parameter independence.

Spearman correlation analysis (Figure 16) reinforced these findings, with FFR-S demonstrating strong positive correlation with BHP-CTD (ρ = 0.86). Other variables exhibited considerably lower correlations, confirming the dominant influence of surface flow rate on bottomhole pressure dynamics during nitrogen lifting operations.

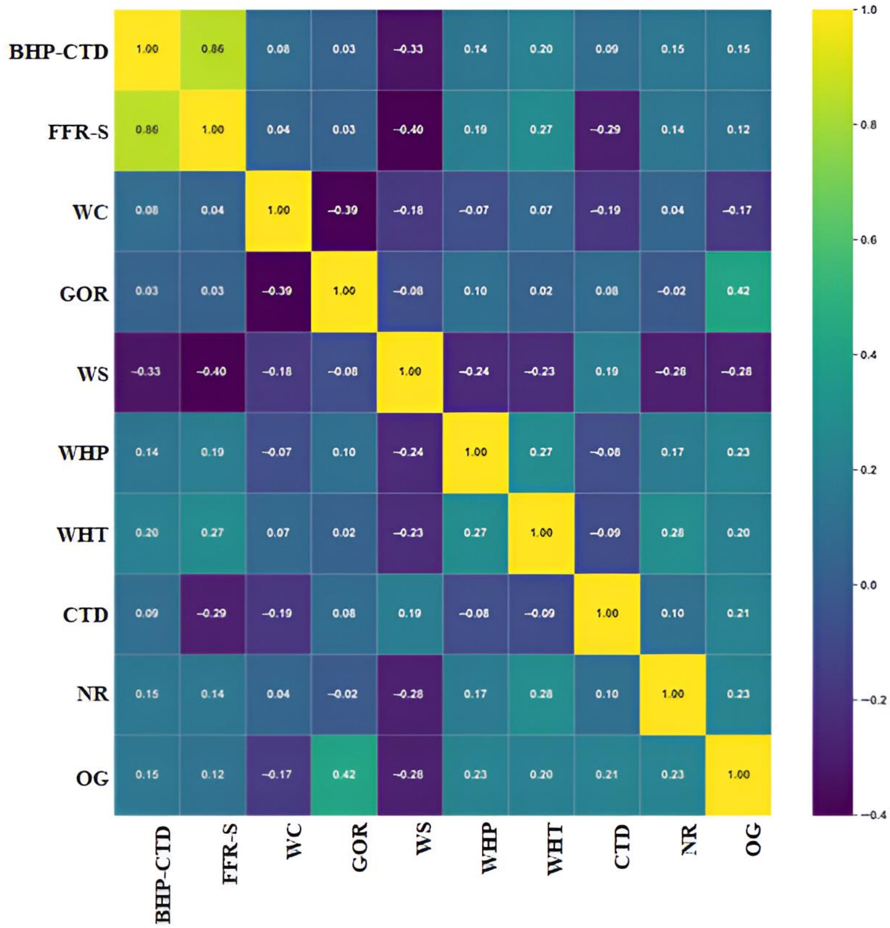

Figure 16. Spearman rank correlation coefficient heatmap displaying monotonic relationships between features in the nitrogen-lifted wells dataset. Fluid flow rate at surface maintains strong positive correlation with bottomhole pressure at coiled tubing depth (ρ = 0.86), confirming its dominant influence on BHP-CTD prediction, while remaining operational parameters show considerably lower correlations with the target variable.

3.4 Data Preprocessing

Data preprocessing is an essential step in the development of machine learning models, as it is a way to guarantee the quality of the information provided as input data, consistency, and proper formatting to ensure optimal algorithm performance. Such systematic reform of raw field data has a direct impact on the model accuracy, the ability to generalize, and the computational efficiency. The preprocessing methodology used in all three applications of hydraulic fracturing, gas lift wells, and nitrogen-lifted wells was a conventional approach of three steps that included data quality evaluation and cleaning, normalization, and strategic division.

3.4.1 Data Quality Assessment and Cleaning

Data quality assessment is the operational stage in preprocessing, which corrects intrinsic inconsistencies and inaccuracies of field-measured data. Raw operations data can easily have missing values, measurement errors, and outliers caused by the sensor failures, communication issues, or severe operational events. The cleaning procedure was initiated by doing systematic identification and treatment of the missing values using two countermeasures: removing the records with incomplete data and extrapolating the data points with the neighbors where possible. This twofold approach balanced the quality assurance and data retention.

Box plot analysis together with domain knowledge was used to detect an outlier, which was used to identify the occurrence of actual extreme operating conditions and incorrect measurements. It was necessary to have this petroleum engineering viewpoint because some of the seemingly extreme values were real working cases instead of the data being corrupted. Elimination of known outliers ensured that model training was not distorted; such outliers can have a disproportionate effect on gradient-based optimization algorithms and distance-based methods. After cleaning operations, data integration was used to combine information that belonged to several different sources into coherent, analyzable datasets.

3.4.2 Data Normalization

Normalization brings features into a common scale and does not allow those features that have large numerical ranges to dominate the learning process, which is essential to algorithms that are sensitive to the magnitudes of features. Variables with thousands of PSI would overwhelm more important parameters with single-digit values without normalization, which adds bias to the gradient descent optimization problem and prevents convergence.

The min-max scaling method, as in Equation (6), was chosen because it maintains the original data correlation and bounds the values to a range of [0, 1]. This method is especially beneficial with neural networks that use sigmoid or tanh activation functions, k-nearest neighbor models that make use of distance measures, and support vector machines that demand uniform scaling of the feature space. The min-max scaling preserves the proportional relationship between data points; it is interpretable within constrained ranges while having the ability to fit non-Gaussian distributions that are typical of operational petroleum engineering scale.

$$Y = \frac{X - A}{B - A} \tag{6}$$

Where

- X is the original (raw) value,
- A is the minimum value in the dataset,
- B is the maximum value in the dataset,
- Y is the normalized value after scaling to the range [0, 1].

The normalization process is used to increase the efficiency of training since the gradient scales become standardized among features, convergence rates are increased, and the sensitivity to initial weights is diminished. It also allows model interpretability, finding proportional feature contributions, which allows easier analysis of prediction mechanisms and decision-making processes.

3.4.3 Data Partitioning Strategy

The last preprocessing step consisted of the systematic splitting of normalized datasets into the training and testing subsets with an 80/20 proportion. With this partitioning strategy, there is enough data to learn the models thoroughly and enough independent samples to assess the performance of the model objectively. The 80% training allows the algorithms to be trained on the complex patterns and relationships in the data, whereas the 20% testing set gives necessary confirmation about the ability to generalize to unknown situations.

The specified method of partitioning directly tackles the issue of overfitting because it keeps training and testing data completely separate, which means that the performance measure will be based on the actual capacities of prediction and not the memorization of training cases. The similarity in the use of the same strategy in all three applications helped to make significant cross-domain comparisons and support the soundness of the devised methodologies.

3.5 Machine Learning Model Development

3.5.1 Model Selection Rationale

3.5.1.1 Diversity of Algorithmic Approaches

The principle of algorithmic diversity was employed to select machine learning models for predicting bottomhole pressure across the three application areas. The use of a single predictive model is not adequate based on the heterogeneous nature of the real data, which have varying levels of nonlinearity, noise, and interactions among features. The underlying assumptions, learning mechanisms, and inductive biases that are represented by different machine learning algorithms imply that some architectures are better suited to certain data distributions, complexities, or operational conditions. Such diversity provides the solution space to be thoroughly explored; the best predictive frameworks are found for each particular application.

The model portfolio includes the full range of regression methods, including simple linear models and complicated ensemble and deep learning models. Linear models can be used to establish marks of baseline performance and interpretability, whereas tree-based approaches can discover nonlinear associations among variables by using hierarchical decision rules. Ensemble methods use the collective knowledge of many weak learners to obtain better predictive power, and neural networks are well-known for their ability to learn interactions that are very complex and multidimensional. The instance-based learning provides non-parametric options, which adapt at the local level to the trends of the data, and the support vector machines provide the robust option in the high-dimensional spaces of features. Lastly, symbolic regression is an intermediate between black-box prediction and physical interpretability as it derives explicit mathematical equations.

3.5.1.2 Comparative Evaluation Strategy

The comparative evaluation approach that has been adopted in all three applications allows strict and unbiased evaluation of the performance of the models working with the same conditions. Through training and testing each algorithm on the same preprocessed datasets, consistent validation methods, and measuring performance of the algorithm by standardized metrics, the research creates a level of fair comparisons that reveal the strengths and weaknesses of each method in a relative manner. This systematic benchmarking can give empirical support in making model selection decisions and can determine model qualities in algorithms that are most beneficial in predicting bottomhole pressure under varying operating conditions.

The assessment system enables us to find trade-offs between competing goals of predictive accuracy, computational efficiency, interpretability, and generalization ability. Although one model might be slightly more accurate on the training data, there are other models that might prove to be much more robust when facing unknown well conditions. The comparative methodology makes sure that the choice of models is not only based on the training performance, but it is also more holistic with respect to the application of the model in the field, with real-time computational requirements, ease of implementation, and maintenance needs.

3.5.1.3 Coverage of Model Complexity Spectrum

The chosen algorithms cover the spectrum of complexity, starting with an easy-to-understand linear model and all the way to complex deep learning architectures, covering every possible method of solution. Simple models have computational efficiency and straightforwardness to interpret, and they are resistant to overfitting but can underfit complicated nonlinear relationships. On the other hand, complex models may describe complex patterns and interactions; however, with the risk of overfitting, large computational resources, and usually acting as black boxes. Through the deployment of models in this spectrum, the study finds the best level of complexity needed to give exact results on bottomhole pressure prediction in any application area.

Such spectrum coverage is especially useful in oil and gas applications, where accuracy versus interpretability is of great importance to the operation. Whereas engineers would want to see interpretable models to explain physical relationships and troubleshooting when the physical process is not functioning as intended, they want to have precise predictions when they need to make key decisions. Symbolic regression in particular is designed to resolve this tension by coming up with interpretable mathematical equations that are also competitive at predicting.

3.5.2 Regression Algorithms Implemented

The complete model of regression algorithms used in the three applications includes sixteen different model types, although not every model was used in each area. Ten models were used in the hydraulic fracturing application, fifteen models in the gas lift wells, and the full set of sixteen algorithms in the study of nitrogen-lifted wells. This difference is prone to the changing methodology during the research process and the inclusion of new algorithmic procedures to expand the prediction potential.

3.5.2.1 Linear Models

Linear regression (LR) is the original regression method, which determines linear correlations between the features of the input and the target variable as a result of least squares maximization. Linear regression, despite its simplicity, offers useful baseline performance measures and is completely interpretable with explicit coefficient values. It is especially appropriate in early exploratory analysis and in those cases where explainability of the model is important, as opposed to maximum predictive accuracy.

Stochastic Gradient Descent (SGD) is based on the principle of linear regression, which works via an iterative optimization algorithm, whereby model parameters are changed according to the randomly selected data subsets instead of the entire dataset. This method has a radical decrease in the computational needs of big datasets and convergence to optimal solutions. The stochastic dynamics provide these positive regularization effects, which might help generalize better, so that SGD is especially useful in high-dimensional cases or streaming data like in the real-world field of use.

3.5.2.2 Tree-Based Models

Decision Trees (DT) build hierarchical decision rules by a recursive binary division of the feature space, attempting to build intuitive tree structures, which reflect the way humans make decisions. Internal nodes are a test of a particular feature, and each branch is a test result; the prediction value is found in each leaf node. Decision trees provide a natural way to deal with both categorical and numerical data, and minimal data preparation is required, as well as the process of decision logic is easily visualized. The individual trees, however, tend to be highly variable and prone to overfitting, which encourages the use of ensemble strategies.

Random Forest (RF) solves decision tree issues by ensemble learning by training a series of decision trees with bootstrapped data samples on randomized feature subsets at each split. The last prediction is a summation of the individual tree predictions at the expense of variance, at a low bias level. Random forests are highly resistant to overfitting, can effectively deal with high-dimensional data, and automatically offer metrics of feature importance. The algorithm is especially effective when it comes to oil and gas applications where there are complex and nonlinear relationships between many operational parameters.

3.5.2.3 Ensemble Boosting Methods

Sequential ensemble learning Gradient Boosting (scikit-learn) (GB-SKL) is an implementation of gradient boosting in which each new weak learner aims to correct the previous ensemble. In contrast to bagging approaches that model each component separately, a gradient boosting approach consists of models that are built on top of each other, and each new model is specifically focused on resolving the leftover errors. The method is usually more accurate than the random forests, but hyperparameters need to be carefully adjusted to avoid overfitting. The scikit-learn implementation is a strong and well-validated algorithm, which can be applied to scientific research.

An alternative boosting strategy, Adaptive Boosting (AdaBoost, ADAB), is a boosting algorithm that modifies the sample weights over sample prediction errors. The cases that are misclassified give more weight in the later iterations, and the algorithm must concentrate on the hard cases. Although it was originally formulated in the classification form, the regression form is useful when having continuous target variables, such as bottomhole pressure. AdaBoost has a specific ability to work through noisy data and to detect some hidden patterns that simple models can overlook.

Extreme Gradient Boosting (XGBoost) is a continuation of the standard gradient boosting with advanced regularization models, parallel processing, and streamlined tree-building models. These further optimize the accuracy and computing efficiency of the algorithm dramatically, and XGBoost is among the most successful machine learning algorithms in the context of competitive benchmarking. The implementation provides inbuilt support of missing values and cross-validation as well as early termination controls against overfitting. XGBoost has proven itself to be very effective in various petroleum engineering problems.

Extreme Gradient Boosting Random Forest (XGBoost-RF) is an experimental model that integrates XGBoost with random forest in terms of both its implementation and its bagging approach, in which trees are trained on random subsamples instead of training on residuals. This mixed method tends to create more generalization compared to pure boosting and still benefits from the computational efficiency of the XGBoost framework. The algorithm is especially useful when the training data are highly noisy or model variance reduction is of more importance than bias minimization.

CatBoost (GB-CB) is a specialized version of gradient boosting that is optimized when using categorical features. The algorithm has ordered boosting and other new categorical encoding schemes that minimize overfitting and maximize prediction accuracy. Although in the bottomhole pressure prediction tasks, the applications mostly relate to numerical characteristics, the smart regularization and powerful default hyperparameters of CatBoost frequently demonstrate good results to the continuous ones, as well. The implementation offers great out-of-the-box performance without much tuning needs.

3.5.2.4 Instance-Based Learning

Uniform Weighting k-Nearest Neighbors (kNN-U): kNN-U is a non-parametric regression method that averages target values of the k nearest training exercises in feature space. The uniform designation means that all the k neighbors are equally weighted irrespective of their distance to the query point. This method does not necessitate a training phase, but rather it stores the complete training data set to compare these data at prediction time. The algorithm is by default sensitive to local patterns of data and can deal with arbitrary decision boundaries, but it is more expensive to compute as the data size and dimension size grow.

The k-Nearest Neighbors distance-weighted (kNN-D) is a modification of the standard algorithm, which gives neighbors a weight based on how far they are from the query point. Neighbors within the likelihood idea have a stronger impact on the prediction compared to those who are farther away and therefore offer more detailed interpolation, which is more representative of the local data structure. The weighting scheme is usually more accurate in areas where the density of data is uneven or where target values vary quickly in the feature space. The distance-weighted form is particularly useful in petroleum engineering, where the relationship between the operational parameters has high locality effects.

3.5.2.5 Support Vector Machines

Support Vector Regression (SVR) is an extension of the principles of the support vector machine classification to continuous target prediction. The algorithm forms an optimal hyperplane in the feature space of high-dimensional space, maximizing the margin with a tolerance of a given error rate of prediction (epsilon). SVMs are made to reach nonlinear relationships through the use of kernel functions that implicitly project data to higher-dimensional data regions that can be separated linearly. The support vector machines show specific capabilities in high-dimensional tasks that require a limited amount of training data, but the computational needs may be prohibitive with very large data sets. The fact that the method has strong theoretical grounds and generalization abilities indicates that it is a good comparative benchmark method.

3.5.2.6 Neural Networks

A Neural Network with L-BFGS Optimizer (NN-LBFGS) is a multilayer perceptron regression model that is based on the Limited-memory Broyden-Fletcher-Goldfarb-Shannon optimization algorithm. This is a generalization of the Newton's method that uses approximations of the inverse Hessian matrix to get a faster convergence rate than the original gradient descent, especially on the smaller to medium-sized dataset. The L-BFGS optimizer is particularly effective in cases where training based on full batches is possible and where the objective function is smooth and well-behaved. Neural networks are very useful in the non-linear interaction between features, which are captured by hierarchical representations taught through the multiple hidden layers.

"Neural Network with Adam Optimizer" (NN-Adam) refers to the use of the Adaptive Moment Estimation optimization algorithm that is an amalgamation of momentum-based gradient descent and adaptive parameter learning rates. Adam has become the default optimizer for a variety of deep learning tasks with little hyperparameter optimization. The algorithm keeps running averages of both the gradients and the squared gradients to be able to navigate through complex loss landscapes easily. This optimizer is especially beneficial in case of noisy gradients or sparse data, which is presented in field measurements.

Neural Network with SGD Optimizer (NN-SGD) is a neural network training application that uses the standard stochastic gradient descent and changes the weights on the basis of the mini-batches instead of the entire dataset. Although it is conceptually simpler than modern optimizers, with a sufficient learning rate schedule, SGD can also compete on a level with current optimizers, especially with such extra terms as momentum or Nesterov acceleration. The simplicity of implementation and convergence properties of the optimizer can be useful in setting performance baselines and in the analysis of optimization dynamics.

The neural networks employed in all three optimizers consisted of a single hidden layer with 50 neurons, utilizing the Rectified Linear Unit (ReLU) activation function. This structure has been obtained after the initial testing because it offers the best balance between model capacity, training stability, and computational efficiency. Compared to the large complex architecture, the relatively compact architecture minimizes the risk of overfitting and is still complex enough to fit the nonlinear relationships of interest. Weight initialization was done in accordance with the Xavier/Glorot scheme, which encourages stable gradient flow in the training process.

3.5.2.7 Symbolic Regression

Genetic Programming-based Symbolic Regression (GP-SR) is a radically new method of machine learning that tries to find the explicit mathematical equations describing the relationships in data automatically. Symbolic regression does not optimize parameters in predefined functional forms but instead searches the space of potential mathematical expressions by evolutionary algorithms based on natural selection. The candidate equations are the subject of repeated iterative evaluation, selection, crossover, and mutation to develop an expression that increasingly well represents the desired relationships between factors whilst also being interpretable through its structural simplicity.

The first benefit of symbolic regression compared to the classical machine learning models is that it does not generate black-box predictions but instead generates mathematical equations that are easy to read. Such equations can give straightforward physical understanding of relations of governance, can be executed by domain experts, and can be implemented in existing software systems without machine learning requirements. The method is especially useful in petroleum engineering practice, where interpretable models are often required by physical insights as well as regulatory compliance. The PySR Python library provided the genetic programming framework, which allows exploration of the expression space efficiently and with a reasonable computational requirement.

3.5.3 Algorithm Descriptions and Theoretical Foundations

3.5.3.1 Linear Model Theory

Linear regression models assume that the target variable can be expressed as a linear combination of input features plus normally distributed error terms. Mathematically, this relationship takes the form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$, where y represents the target variable (BHP), x_1 through x_n denote input features, β_0 through β_n are coefficients to be estimated, and ε captures random error. Ordinary least squares estimation determines coefficient values by minimizing the sum of squared residuals between predicted and actual target values.

The method's theoretical foundations rest on several key assumptions: linearity of relationships, independence of observations, homoscedasticity (constant error variance), and normal distribution of residuals. When these assumptions hold reasonably well, linear regression provides unbiased, minimum-variance estimates with well-characterized uncertainty. Violations of these assumptions motivate regularization techniques such as Ridge regression (L2 penalty) and Lasso regression (L1 penalty), which constrain coefficient magnitudes to improve generalization. Elastic Net regularization, implemented in several models, combines both L1 and L2 penalties, balancing feature selection capabilities with coefficient shrinkage.

3.5.3.2 Tree-Based Algorithms Principles

Tree-based algorithms repeatedly divide feature space into subsets that are small binary splits and which maximize information gain or minimize prediction error in the resulting subsets. Each node of the algorithm considers all potential points of division based on all features and chooses the division that best divides the data based on a given criterion. In the case of regression work, it is customary to measure the reduction of mean squared error or the reduction of variance. This greedy partitioning process goes on until the process stops because of satisfying stopping criteria, like attaining maximum tree depth, minimum node size, or having a sufficiently small error reduction.

Single decision trees have a reasonably intuitive interpretation and need few data preprocessing steps but overfit the training data, modeling noise instead of any generalizable trend. Bootstrap aggregating (bagging) is a technique used by the random forests to overcome this weakness by training many trees on random data samples and averaging their predictions. Further decorrelation of individual trees by the extra randomization of feature selection at each split decreases the variance of the ensemble. Gradient boosting is a different method, which is used to chain-train trees and correct the errors of earlier ones, and is generally more accurate but must be more carefully regularized to avoid overfitting.

3.5.3.3 Ensemble Learning Mechanisms

Ensemble learning is a combination of predictions made by different base models in order to perform better than any other single model. The basic idea in ensemble success is that different models can have different errors on different cases, and averaging these predictions eliminates errors made by individual models but retains errors made by other models. Ensemble diversity may be attained by training on varying data subsets (bagging), concentrating on varying facets of the issue (boosting), or by training completely dissimilar algorithms (stacking).

The algorithms of boosting, in particular, directly apply sequential error correction, in which every new model is explicitly focused on those cases, which are not well predicted by the existing ensemble. AdaBoost achieves this by performing adaptive instance weighting with initial model emphasis on those cases that it misclassifies; thus, the next models it builds concentrate on such hard cases. Gradient boosting is a continuation of this idea, but new models are trained such that they predict the residual errors (negative gradients of the loss function) of the current ensemble predictions. This method can be used to optimize arbitrary differentiable loss functions in a sophisticated way and is generally able to reach state-of-the-art accuracy in a variety of applications.

Contemporary gradient boosting algorithms have a number of refinements such as regularization penalties on tree complexity, stochastic sampling of training examples and features, and early termination on validation set performance. These do not only increase the generalization in a miraculous way but also enhance computational efficiency with the help of optimized algorithms and parallel processing.

3.5.3.4 Instance-Based Learning Concepts

Instance-based learning, including k-nearest neighbors algorithms, is another paradigm that is radically different from model-based ones. Instead of establishing explicit and functional relationships through the training process, instance-based approaches merely archive training examples and postpone computations to the prediction time. To obtain a new query point, the algorithm finds the k nearest training points, according to a distance measure (usually the Euclidean distance in normalized feature space), and sums together the target values of these points to form a prediction.

The k parameter balances between bias and variance: lower values have flexible and low-bias predictions that trace the local data but have large variance and are sensitive to noise. Greater k values average more widely, minimizing variance, but can also add bias, adding instances that are not similar. Distance weighting offers further refinement, reducing the effect of more distantly located neighbors, and allows adaptive interpolation, which gives more respect to local data structure.

The instance-based process is intrinsically able to deal with arbitrary decision boundaries and local data characteristics, not necessarily specifying their models explicitly. But they experience serious problems in high-dimensional spaces (the curse of dimensionality) in which distance measures are less meaningful, and the cost of computation is linear in training set size as each prediction involves distance computation to all the stored examples.

3.5.3.5 Support Vector Machines Framework

SVMs build the best dividing hyperplanes in multidimensional feature areas to maximize the distance between varying classes or the distance between the prediction function and the training samples (in a regression scenario). In the case of regression problems, SVMs use an ε-insensitive loss term that ignores all errors that are less than a given threshold ε and concentrates model capacity on cases with higher prediction errors. The strategy encourages sparse solutions in which predictions are restricted to a small set of training examples (support vectors) that are close to the decision boundary or edges of prediction tubes.

The kernel trick allows SVMs to draw large nonlinear relationships by implicitly reconfiguring the data to higher-dimensional spaces in which linear division can be made. The most widespread kernel functions are linear, radical basis function (RBF), and sigmoid kernels, which induce a variety of feature space geometries. The RBF kernel is especially popular because it fits smooth and nonlinear relationships with the number of hyperparameters that are relatively small.

Support vector machines have solid theoretical underpinnings in the statistical learning theory and offer generalization guarantees on the basis of margin maximization principles. They are useful in most petroleum engineering applications because of their ability to perform well in high-dimensional environments using little training data. Computational demands are, however, not scaling well to data sets that are much larger in size, and the optimal choice of kernel can demand a lot of experimentation.

3.5.3.6 Neural Network Architecture

Artificial neural networks arrange processing units (neurons) in interlinked layers, with each neuron calculating a weighted sum of the inputs and then acting on them by a nonlinear activation function. The multilayer perceptron architecture used in this study consists of an input layer that takes the values of the features, one or more layers that transform the input and give the most nonlinear transformation, and finally an output layer that makes predictions. The prediction operation follows the network by using information forward, and training uses backpropagation to find the gradients of the loss function with respect to all weights.

The activation function used has a large effect on the training dynamics and behavior of networks. The default activation of hidden layers has become the Rectified Linear Unit (ReLU) $f(x) = \max(0, x)$, which is computationally fast, reduces the issue of vanishing gradients, and encourages sparse activations. The unlimited continuous predictions that can be made by the output layer are based on the linear activation in the output layer that is appropriate for regression problems.

The model capacity and expressiveness are regulated by network depth (number of layers) and width (neurons per layer). Finer networks can learn hierarchical representations of features of more and more abstraction, and broader networks gain capacity in single layers. A single hidden layer design with 50 neurons used in this study offers enough capacity to learn nonlinear relationships with complexity while having the advantage of being computationally efficient and avoiding overfitting as compared to multiple layers.

The neural networks' training also needs a proper choice of the optimization algorithms. L-BFGS estimates the Hessian matrix in order to reach a faster convergence when dealing with small datasets whose loss functions are smooth. Adam scales the learning rate of individual parameters depending on the statistics of the gradients, which offers superior performance in a wide range of problems.

3.5.3.7 Symbolic Regression Methodology

In symbolic regression, genetic programming is used to simulate natural selection in populations of mathematical expressions by evolving them with genetic programming. Every expression is a tree with the internal nodes symbolizing mathematical functions (addition, multiplication, trigonometric functions, etc.) and the leaf nodes symbolizing either input variables or numerical constants. The evolutionary process starts with the initial population, which has been randomly generated, and continues its course in a series of repetitions of fitness assessment, selection of parents, application of genetic operators (crossover and mutation), and replacing the population.

Fitness evaluation is an assessment of the accuracy of each candidate expression to training data, which is usually assessed by mean squared error or other regression-related measures. The mechanisms of selection prefer those individuals that are fittest to reproduce and ensure diversity by using stochastic processes. Crossover exchanges overlay parent expression fragments and recombines favorable substructures, whereas mutation randomly alters and examines new expression structures. Other operators such as simplification are useful to decrease the expression complexity to encourage interpretable solutions.

The main problem with symbolic regression is how to balance accuracy and complexity, complex expressions have an inherent advantage of fitting the training data better but have the disadvantage of being less interpretable and causing overfitting. Multi-objective optimization methods directly express accuracy and complexity as conflicting goals and retain a Pareto front of non-dominated solutions, which provide trade-offs of various types. The implementation of PySR utilizes advanced methods such as constant optimization using simulated annealing, physical consistency using dimensional analysis, and better exploration using parallel island population.

Symbolic regression generates explicit mathematical equations that can be validated by domain experts on the basis of physical principles, implemented in current software without machine learning requirements, and analyzed to learn about the relationships that govern it. This interpretability is especially useful when applied to petroleum engineering, where regulatory compliance, operational transparency, and theoretical understanding are of great importance in conjunction with predictive accuracy.

3.5.4 Application-Specific Model Configurations

3.5.4.1 Hydraulic Fracturing Models

Ten machine learning models were used in the hydraulic fracturing application. This selection was the first step of the study and was aimed at defining the baseline performance within the variety of algorithmic families without using more sophisticated versions of boosting or more symbolic regression.

Table 4 is the description of the hyperparameter settings of these models where cautious tuning of predictive accuracy and computational efficiency parameters is made. The learning rate was 0.099, and the maximum depth of the trees was 6, which was used to minimize the use of gradient boosting to encourage stable predictions and reduce the risk of overfitting. Random Forest also used only 10 trees with a minimum subset size of 5, which also focused more on ensemble diversity than on the number of models. Linear kernels were used in support vector machines, as opposed to RBF, and this was due to the fact that the support vector machines emphasized computational efficiency given the large scale of the data (42 wells).

The configuration of the neural network was consistent across optimizers, with 1500 neurons in the hidden layer and ReLU activation and a regularization strength of 0.04. The size of the hidden layer was large, which was associated with the complexity of the dynamics of pressure in hydraulic fracturing and the size of the training dataset. The optimizers used had different numbers of maximum iterations: 500 in L-BFGS and Adam since they achieve convergence faster, whereas SGD has 1000, which is higher to stabilize its slower learning dynamics. These settings were a result of initial experiments of optimization between training time, convergence stability, and predictive performance.

Table 4. Hyperparameter configurations for ten machine learning models implemented in the hydraulic fracturing bottomhole pressure prediction application. Model-specific parameters include tree counts, learning rates, regularization strengths, kernel types, hidden layer dimensions, activation functions, and optimization algorithms, with configurations tuned to balance predictive accuracy and computational efficiency.

Model	Algorithm Parameters
GB	▪ Number of trees is 100. ▪ Learning rate is 0.099. ▪ Limit depth is 6. ▪ Minimum subset size is 2. ▪ The fraction of training instances is 1.
AdaBoost	▪ The number of estimators is 100. ▪ Learning rate is 1. ▪ Regression loss function is linear.
RF	▪ Number of trees in the forest is 10. ▪ Minimum subset size is 5.
SVMs	▪ SVM Cost is 1. ▪ Regression loss epsilon: 0.1. ▪ Kernel type is radial basis function. ▪ Iteration limit is 100.
DT	▪ Minimum instances in leaves are 19. ▪ Minimum subset size: 9. ▪ Maximal tree depth is 100. ▪ The stopping point is at 95% of the majority.
kNN	▪ Number of nearest neighbors is 5. ▪ Metric is Euclidean. ▪ Weights are uniform.
LR	▪ Regularization: Ridge regression (L2) ▪ Alpha: 0.05
NN (L-BFGS)	▪ Neurons per hidden layer are 1500. ▪ Activation is ReLu. ▪ Solver is L-BFGS. ▪ Regularization: 0.04. ▪ Maximum iterations are 500.
NN (Adam)	▪ Neurons per hidden layer are 1500. ▪ Activation is ReLu. ▪ The solver is Adam. ▪ Regularization: 0.04. ▪ Maximum iterations are 500.
SGD	▪ Regression loss function is squared loss. ▪ Regularization is lasso (L1).

	▪ Learning rate is constant. ▪ Initial learning rate is 0.01. ▪ Number of iterations is 1000.

3.5.4.2 Gas Lift Well Models

The gas lift wells application added to the model portfolio fifteen algorithms with other variants of boosting, such as XGBoost, XGBoost Random Forest, and CatBoost, as well as the uniform-weighted k-nearest neighbors and symbolic regression. This growth also signified the development of methodology and the acknowledgment that the complicated multiphase flow interactions in gas transit operations could be better portrayed by higher-level ensemble techniques as compared to conventional algorithms.

Table 5 provides hyperparameter settings that show greater complexities in tuning the model. The XGBoost was used with 200 trees and a learning rate of 0.05, which balances both overfitting and accuracy by heavy regularization. The L2 regularization parameter (lambda) was 1, and the maximum depth was limited to 6. CatBoost also used the same 100 trees and the learning rate of 0.05 and L2 penalty of 3 and the maximum depth of 5, which indicates its strong default settings.

The neural network constructions were moved to smaller ones consisting of 50 hidden neurons, less regularization (0.01), and more maximal iterations (1000 per all optimizers). This modification was due to the smaller dataset than hydraulic fracturing, where the model could easily overfit due to overcapacity. The symbolic regression component proposed evolutionary computing with 100 iterations and an expression size of 20.

Table 5. Hyperparameter specifications for fifteen machine learning and symbolic regression models applied to gas lift well bottomhole pressure prediction. Configurations include expanded

boosting algorithms (XGBoost, XGBoost-RF, CatBoost) with detailed regularization parameters, refined neural network architectures with 50 hidden neurons, and genetic programming settings for symbolic regression (100 iterations, maximum expression size 20).

Model	Model Parameters
GB (scikit-learn)	▪ Number of trees is 100 ▪ Learning rate is 0.1 ▪ Limit depth of individual trees is 3 ▪ Does not split subsets smaller than 10 ▪ The fraction of training instances is 0.8
EGB (xgboost)	▪ Number of trees is 200 ▪ Learning rate is 0.05 ▪ Lambda (L2 regularization) is 1 ▪ Limit depth of individual trees is 6 ▪ Fraction of training instances is 0.8 ▪ Fraction of features for each tree is 0.8 ▪ Fraction of features for each level is 1 ▪ Fraction of features for each split is 0.8
EGB-RF (xgboost)	▪ Number of trees is 200 ▪ Learning rate is 1 ▪ Lambda (L2 regularization) is 2 ▪ Limit depth of individual trees is 5 ▪ Fraction of training instances is 0.8 ▪ Fraction of features for each tree is 0.7 ▪ Fraction of features for each level is 1 ▪ Fraction of features for each split is 0.8
GB (catboost)	▪ Number of trees is 100 ▪ Learning rate is 0.05 ▪ Lambda (L2 regularization) is 3 ▪ Limit depth of individual trees is 5 ▪ Fraction of features for each tree is 0.8
AdaBoost	▪ The number of estimators is 100 ▪ Learning rate is 0.05 ▪ Boosting method is SAMME.R (Real boosting) ▪ Regression loss function is square
RF	▪ Number of trees in the forest is 10 ▪ Number of attributes per split is 5 ▪ Limit depth of individual trees is 5 ▪ Do not split subsets smaller than 5
SVMs	▪ SVM cost is 1 ▪ Regression loss epsilon: 0.1 ▪ Kernel type is linear ▪ Numerical tolerance: 0.001 ▪ Iteration limit is 1000
DT	▪ Minimum instances in leaves are 15 ▪ Does not split subsets smaller than 7

	▪ Maximal tree depth is 10 ▪ The stopping point is at 95% of the majority
kNN (distance)	▪ Number of nearest neighbors is 5 ▪ Metric is Euclidean ▪ Weight is by distance
KNN (uniform)	▪ Number of nearest neighbors is 5 ▪ Metric is Euclidean ▪ Weight is uniform
LR	▪ Fit intercept ▪ Regularization: Elastic Net regression ▪ Alpha: 10 ▪ L1:L2 mixing ratio: 0.5:0.5
NN (L-BFGS)	▪ Sklearn's Multi-layer Perceptron (MLP) algorithm ▪ Neurons per hidden layer are 50 ▪ Activation is ReLu ▪ Solver is L-BFGS-B ▪ Regularization: 0.01 ▪ Maximum iterations are 1000
NN (Adam)	▪ Sklearn's Multi-layer Perceptron (MLP) algorithm ▪ Neurons per hidden layer are 50 ▪ Activation is ReLu ▪ Solver is Adam ▪ Regularization: 0.01 ▪ Maximum iterations are 1000
NN (SGD)	▪ Sklearn's Multi-layer Perceptron (MLP) algorithm ▪ Neurons per hidden layer are 50 ▪ Activation is ReLu ▪ Solver is SGD ▪ Regularization: 0.01 ▪ Maximum iterations are 1000
SGD	▪ Regression loss function is squared loss ▪ Regularization is Elastic Net ▪ Elastic Net mixing ratio is 0.5 ▪ Regularization strength is 0.001 ▪ Learning rate is constant ▪ Number of iterations is 1000
GP-SR	▪ Number of evolutionary iterations is 100 ▪ Maximum expression size is 20

3.5.4.3 Nitrogen-Lifted Wells Models

The nitrogen-lifted wells study used the full range of sixteen models as methodological refinement on the top of the three applications. The model set included all the algorithms used in earlier works, as well as variants, allowing benchmarking of all the models under the toughest conditions: predicting bottomhole pressure at the depth of coiled tubing when performing transient nitrogen lifting operations.

The hyperparameter specifications, given in Table 6, were more or less the same configurations used in the gas lift study with slight alterations. The tree-based ensemble models did not change the parameters they had shown to be effective: 100-200 trees per algorithm, low learning rates (0.05-0.1), low maximum depth (3-6), and good regularization. The 50-neuron-long hidden layers, 0.01 regularization, and 1000 maximum iterations used in neural networks ensured the same capacity across the increased range of algorithms.

The size of the dataset (518 wells) was neither too large (not exceeding the hydraulic fracturing application) nor too small (not too small to implement the balanced hyperparameter selection). The fact that all the sixteen models were successfully run with no computational quantizing or convergence difficulties confirmed the soundness of all the developed settings and gave the most complete comparison of algorithms throughout the research program.

Table 6. Comprehensive hyperparameter configurations for sixteen machine learning models implemented in nitrogen-lifted well bottomhole pressure prediction at coiled tubing depth. The complete algorithm suite encompasses linear models, tree-based methods, ensemble boosting techniques, instance-based learning, support vector machines, neural networks with multiple optimizers (L-BFGS, Adam, SGD), and genetic programming-based symbolic regression.

Model	Hyperparameters
GB-SKL	▪ Total trees: 100 ▪ Learning step size: 0.1 ▪ Maximum tree depth: 3 ▪ Minimum subset size for splitting: 10 ▪ Training sample ratio: 0.8
XGB	▪ Total trees: 200

	▪ Learning step size: 0.05 ▪ L2 regularization (lambda): 1 ▪ Maximum tree depth: 6 ▪ Training sample ratio: 0.8 ▪ Feature fraction per tree: 0.8 ▪ Feature fraction per level: 1.0 ▪ Feature fraction per split: 0.8
XGB-RF	▪ Total trees: 200 ▪ Learning rate: 1 ▪ L2 regularization (lambda): 2 ▪ Maximum depth of trees: 5 ▪ Training sample ratio: 0.8 ▪ Features per tree: 0.7 ▪ Features per level: 1.0 ▪ Features per split: 0.8
GB-CB	▪ Total trees: 100 ▪ Learning rate: 0.05 ▪ L2 penalty: 3 ▪ Maximum depth: 5 ▪ Feature subset for each tree: 0.8
ADAB	▪ Number of estimators: 100 ▪ Learning rate: 0.05 ▪ Boosting type: SAMME.R (Real boosting) ▪ Loss function (regression): squared error
RF	▪ Size of forest: 10 trees ▪ Number of attributes per split: 5 ▪ Maximum depth: 5 ▪ Minimum samples for splitting: 5
SVMs	▪ SVM penalty parameter (C): 1 ▪ Epsilon for regression margin: 0.1 ▪ Kernel: Linear ▪ Tolerance: 0.001 ▪ Max iterations: 1000
DT	▪ Minimum leaf size: 15 ▪ Minimum split size: 7 ▪ Deepest allowable tree: 10 ▪ Stopping threshold: 95% of dominant class
KNN-D	▪ k value: 5 neighbors ▪ Distance metric: Euclidean ▪ Weighting scheme: distance-based
KNN-U	▪ k value: 5 neighbors ▪ Distance metric: Euclidean ▪ Weighting scheme: uniform
LR	▪ Intercept term: included ▪ Regularization type: Elastic Net ▪ Alpha (regularization strength): 10 ▪ L1/L2 mix: 0.5: 0.5

NN-LBFGS	▪ MLP (scikit-learn implementation) ▪ Hidden layer size: 50 neurons ▪ Activation function: ReLU ▪ Optimizer: L-BFGS-B ▪ Regularization weight: 0.01 ▪ Max iterations: 1000
NN-Adam	▪ MLP (scikit-learn implementation) ▪ Hidden layer size: 50 neurons ▪ Activation function: ReLU ▪ Optimizer: Adam ▪ Regularization weight: 0.01 ▪ Max iterations: 1000
NN-SGD	▪ MLP (scikit-learn implementation) ▪ Hidden layer size: 50 neurons ▪ Activation function: ReLU ▪ Optimizer: SGD ▪ Regularization weight: 0.01 ▪ Max iterations: 1000
SGD	▪ Loss function: squared loss ▪ Regularization: Elastic Net ▪ Elastic Net mixing: 0.5 ▪ Regularization weight: 0.001 ▪ Learning rate strategy: constant ▪ Number of training iterations: 1000

Chapter 4: Results and Discussion

4.1 Model Results

4.1.1 Hydraulic Fracturing Application

4.1.1.1 Model Performance Comparison

The data set was preprocessed, and training and test sets were selected out of the data set. I divided the data into two, which consisted of 80% training set and the remaining 20% test set. Evaluation of ten machine learning models to predict bottomhole pressure during hydraulic fracturing operations found that there was significant variation in predictive accuracy. Table 7 shows detailed metrics such as MSE, RMSE, MAE, and R^2 of all the models. The ensemble techniques and the neural networks showed better performance due to their ability to model complex nonlinear relationships of hydraulic fracturing pressure dynamics.

Table 7. Performance metrics for ten ML models: Gradient Boosting (R^2 = 0.931), Decision Trees (R^2 = 0.903), and Neural Network L-BFGS (R^2 = 0.901) demonstrated superior accuracy.

Model	MSE	RMSE	MAE	R^2
Gradient boosting	0.003	0.058	0.039	0.931
Tree	0.004	0.061	0.04	0.903
Neural network (L-BFGS)	0.004	0.061	0.044	0.901
kNN	0.004	0.062	0.04	0.898
Neural network (Adam)	0.004	0.063	0.046	0.896
Random forest	0.004	0.065	0.041	0.89
AdaBoost	0.005	0.069	0.043	0.875
Linear regression	0.006	0.077	0.058	0.846
SGD	0.006	0.079	0.062	0.835
SVM	0.051	0.225	0.184	0.511

Gradient Boosting was the most successful model (R^2 = 0.931, MSE = 0.003); next in line were Decision Trees (R^2 = 0.903, MSE = 0.004) and Neural Network L-BFGS (R^2 = 0.901, MSE = 0.004). Other models such as k-Nearest Neighbors and neural networks using Adam and SGD optimizers performed well, with each having a result above R^2 = 0.89. Conversely, more basic models performed less well: Linear Regression ($R^2 = 0.846$), Stochastic Gradient Descent ($R^2 = 0.835$), and especially Support Vector Machines ($R^2 = 0.511$) to emphasize the weakness of linear assumptions in predicting the behavior of hydraulic fracturing pressure.

4.1.1.2 Best Model Performance and Visualization

The gradient boosting model has been able to perform better because of its sequential error correction feature whereby each additional tree focuses on the residual errors made by the earlier iterations. The correlation between downhole gauges and model predictions of normalized actual bottomhole pressure is plotted in Figure 17. The close grouping of the points on the straight line indicates high accuracy even in the full range of pressure. The blue-yellow gradient of color depicts the same performance in all levels of pressure, and the small amount of scatter shows that there is low prediction bias and variance.

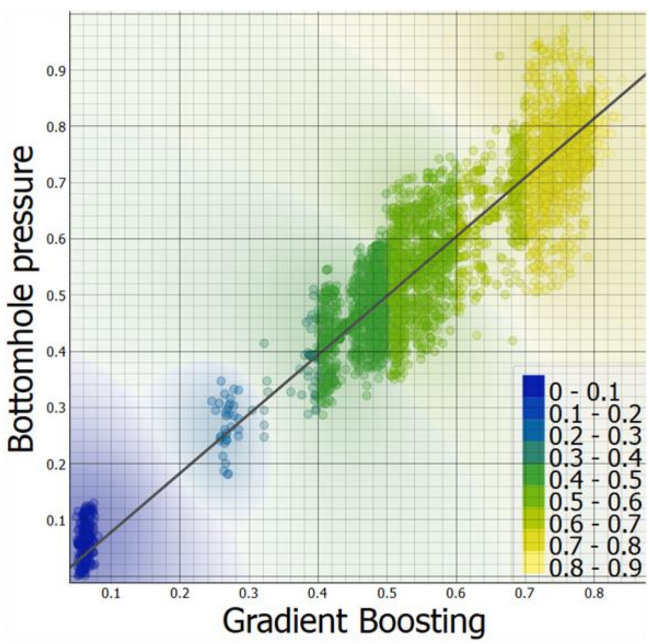

Figure 17. Normalized actual versus predicted BHP by Gradient Boosting model.

4.1.1.3 SHAP-Based Feature Importance Analysis

Figure 18 provides the SHAP analysis, which offers interpretable information on the contribution of features not based on mere correlations. Surface pressure turned out to be the major predictor with the largest range of SHAP value. The positive contributions to predicted bottomhole pressure were always produced by high values of surface pressure (red points), whereas low values (blue points) decreased the prediction, which is consistent with the basic principles of fluid mechanics.

The depth of the pressure gauge exhibited the second-best impact, and the SHAP values showed its contribution to the determination of hydrostatic pressure. Slurry flow rate was found to be moderately affected by complex patterns that are indicative of competing forces of friction losses and pressure transmission. Secondary parameters such as surface proppant concentration, tubing inside diameter, gel load, proppant size, and specific gravity had a narrower range of SHAP but made significant contributions by interacting with other factors.

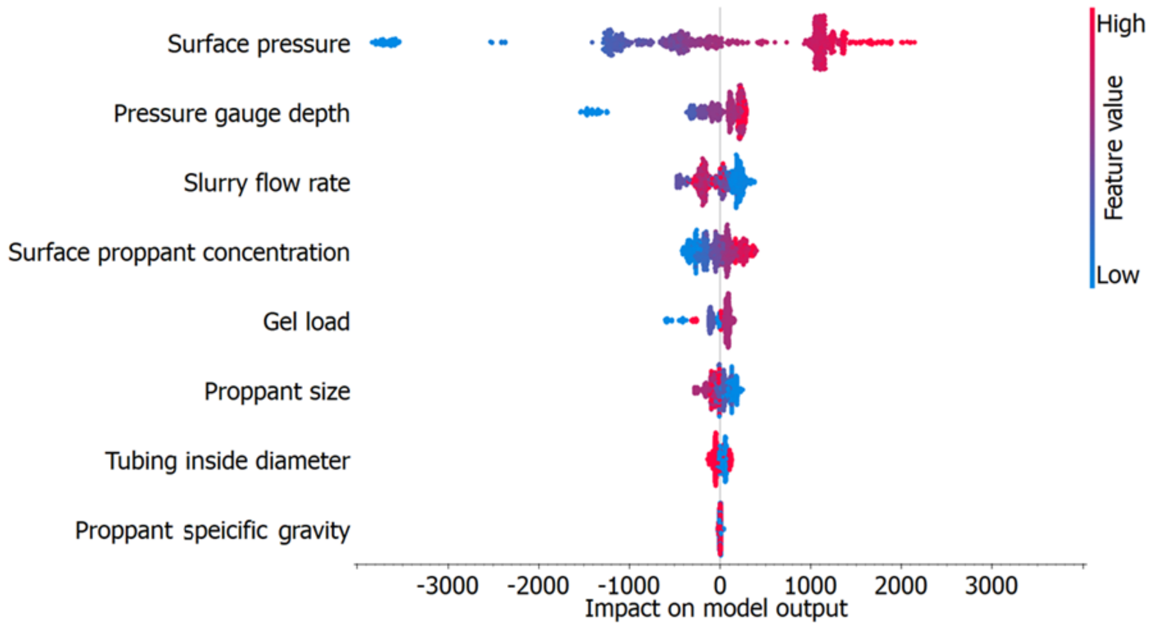

Figure 18. SHAP analysis revealing surface pressure as dominant predictor, followed by pressure gauge depth and slurry flow rate.

4.1.2 Gas Lift Well Application

4.1.2.1 Model Performance Comparison

The data set was preprocessed, and training and test sets were selected out of the data set. I divided the data into two, which consisted of 80% training set and the remaining 20% as the test set. The enlarged model range of fifteen models showed a great deal of performance dispersion, as reported in the Figure 19 performance heatmap. The neural network using the L-BFGS optimizer gave the best results (R^2=0.965, MSE=0.001, RMSE=0.037, MAE=0.021), and so it can be considered the most accurate in this case. The quasi-Newton method of the L-BFGS optimizer was especially successful on the moderate-sized dataset (3,648 points of 304 wells).

Higher-order ensemble models such as XGBoost, CatBoost, and variants of Gradient Boosting gave competitive performance with an average of over $R^2 = 0.82$. The instance-based approaches and support vector machines demonstrated fair but worse results, whereas linear models failed to deal with the nonlinear nature of pressure-flow relationships.

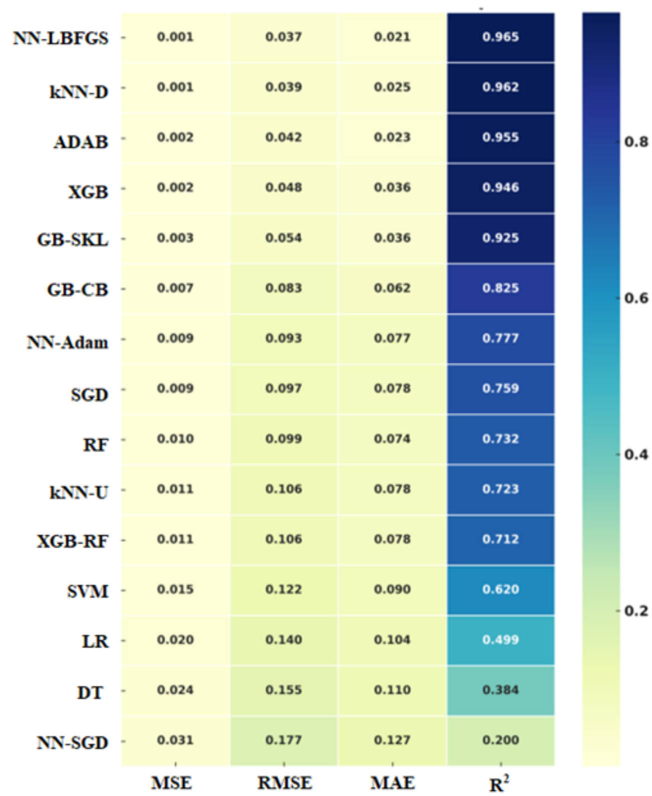

Figure 19. Performance heatmap for fifteen models: Neural Network L-BFGS achieved $R^2 = 0.965$, RMSE = 0.037.

4.1.2.2 Best Model Performance and Visualization

Figure 20 shows the comparison of the normalized actual bottomhole pressure measurements and the neural network L-BFGS predictions. The outstanding match along the diagonal and the transition of color gradient prove the precision throughout the entire range of operation. The 50-neuron hidden layer using ReLU activation was adequate to represent the complicated correlations among eleven parameters as inputs without incurring excessive computational costs.

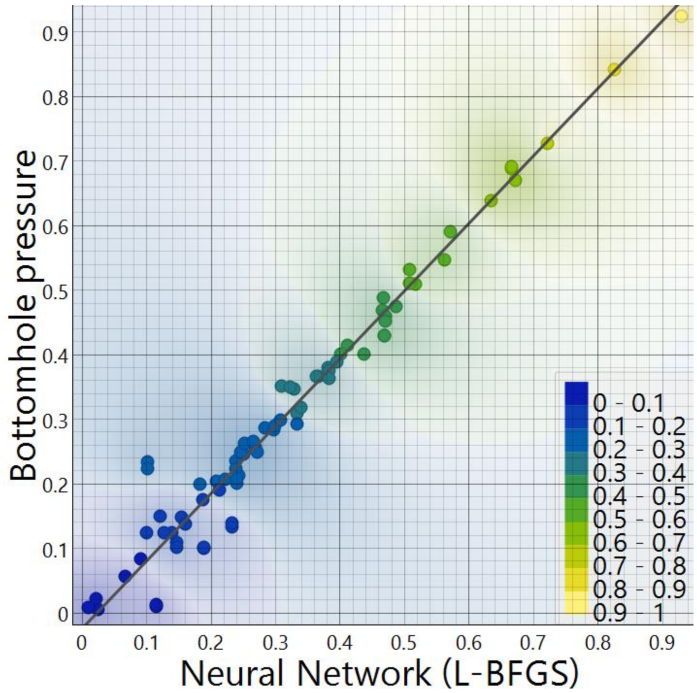

Figure 20. Normalized actual versus predicted BHP showing exceptional alignment ($R^2 = 0.965$) for Neural Network L-BFGS.

4.1.2.3 SHAP Analysis and Symbolic Regression

Figure 21 indicates different patterns of feature importance as compared to correlation-based rankings. The depth of injection point turned out to be the most essential and showed the largest distribution of SHAP. The depth of tubing and the rate of fluid flow had a secondary yet significant effect with complex non-monotonic patterns, which showed indication of an interaction effect.

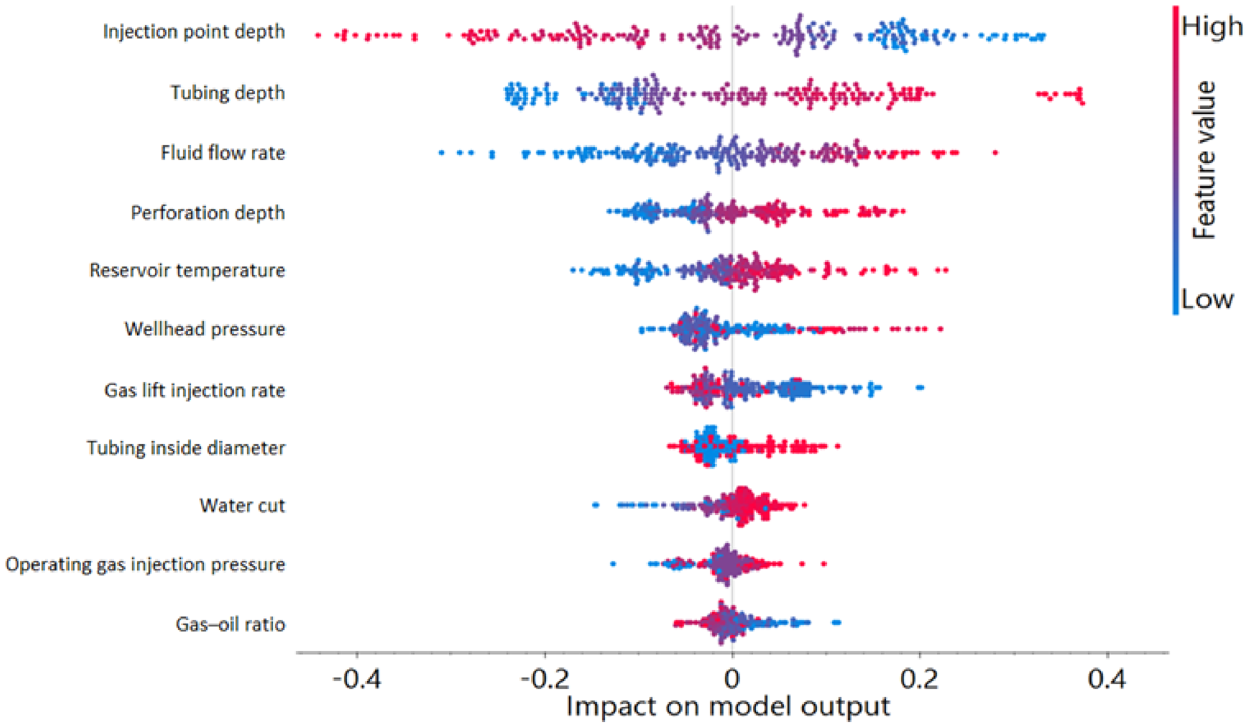

Figure 21. SHAP analysis identifying injection point depth, tubing depth, and fluid flow rate as primary predictors.

The symbolic regression produced via genetic programming produced many possible equations, as seen in Table 8, ranked by complexity and loss values. Equation (7) was chosen as the best with $R^2 = 0.802$, MSE = 0.007, RMSE = 0.085, MAE = 0.068, complexity = 15, and loss = 0.010488. This formula takes into account water cut, depth of injection point, depth of tubing, rate of fluid flow, and their interrelations using square rooting and trigonometric operations, which gives an understandable and implementable version without machine learning requirements.

The variables in the symbolic regression equations below are in dimensionless form and have been normalized to the range [0, 1] using the min-max normalization constants in Section 3.4.2. Inputs to these equations need to be normalized and the output denormalized to compute bottomhole pressure in engineering units.

$$\left(\sqrt{\mathrm{FFR}} \times 0.39444104 + (\mathrm{TD} - (\mathrm{IPD} \times \cos(\mathrm{WHP})))\right) \times (\mathrm{WC} + \mathrm{IPD}) \tag{7}$$

Table 8. Symbolic regression candidates ordered by complexity, with Equation (4) selected as optimal balance.

Complexity	Loss	Equation
1	0.040771	0.3271417
3	0.028801	WC/2.23883
4	0.028639	Sin (WC) × 0.51090336
5	0.024993	$\frac{\mathrm{WC}}{e^{\cos(\mathrm{PD})}}$
6	0.024519	WC/(cos (PD)/0.36379465)
7	0.019756	WC × (TD − IPD + 0.42867288)
8	0.016829	PD − ((WC × −0.37555227) + sin (IPD))
9	0.015375	(OGIP + WC) × (0.2983757 − (IPD − TD))
11	0.014491	(WC + OGIP) × (0.24029268 − (IPD − (PD × 1.095866)))
13	0.013551	((((FFR × 0.2097214) + 0.19042374) × (TD + WC)) + TD) − IPD
14	0.012628	(TD − ((FFR × WC) × −0.55305517)) − (cos (WHP) × (IPD/1.3678912))
15	0.010488	$\left(\sqrt{\mathrm{FFR}} \times 0.39444104 + (\mathrm{TD} - (\mathrm{IPD} \times \cos(\mathrm{WHP})))\right) \times (\mathrm{WC} + \mathrm{IPD})$
16	0.010348	(((WC + TD) × ((FFR × 0.2841561) + 0.13601056)) + TD) − (IPD × cos (WHP))
17	0.009773	$\left(\frac{\sqrt{\mathrm{FFR} \times 0.27031562} + (\mathrm{TD} - (\cos(\mathrm{WHP}) \times \mathrm{IPD}))}{1.3074819}\right) \times (\mathrm{WC} + \mathrm{TD})$
18	0.00901	$\left(\frac{\mathrm{TD} - (\cos(\sqrt{\mathrm{WHP}}) \times \mathrm{IPD})}{1.3286747} + \sqrt{\mathrm{FFR} \times 0.12889326}\right) \times (\mathrm{WC} + \mathrm{PD})$
19	0.008182	$(\mathrm{WC} + \mathrm{TD}) \times \left(\frac{\mathrm{TD} - (\cos(\sqrt{\mathrm{WHP}}) \times \mathrm{IPD})}{e^{\mathrm{GLIR}}} + \sqrt{\mathrm{FFR} \times 0.13287625}\right)$

4.1.3 Nitrogen-Lifted Wells Application

4.1.3.1 Model Performance Comparison

The data set was pre-processed, and training and test sets were selected out of the data set. I divided the data into two, which consisted of 80% training set and the remaining 20% test set. The thorough analysis of the sixteen models showed performance patterns in line with former applications. The heatmap (Figure 22) shows the metrics of all the algorithms. Neural Network L-BFGS had the best performance (R^2 = 0.987, MSE = 0.001, RMSE = 0.014, MAE = 0.011).

Ensemble boosting strategies also performed well, with XGBoost, CatBoost, and Gradient Boosting all surpassing R^2 = 0.96. Linear models demonstrated typical shortcomings, as linear regression was not doing so well at R^2 = 0.001.

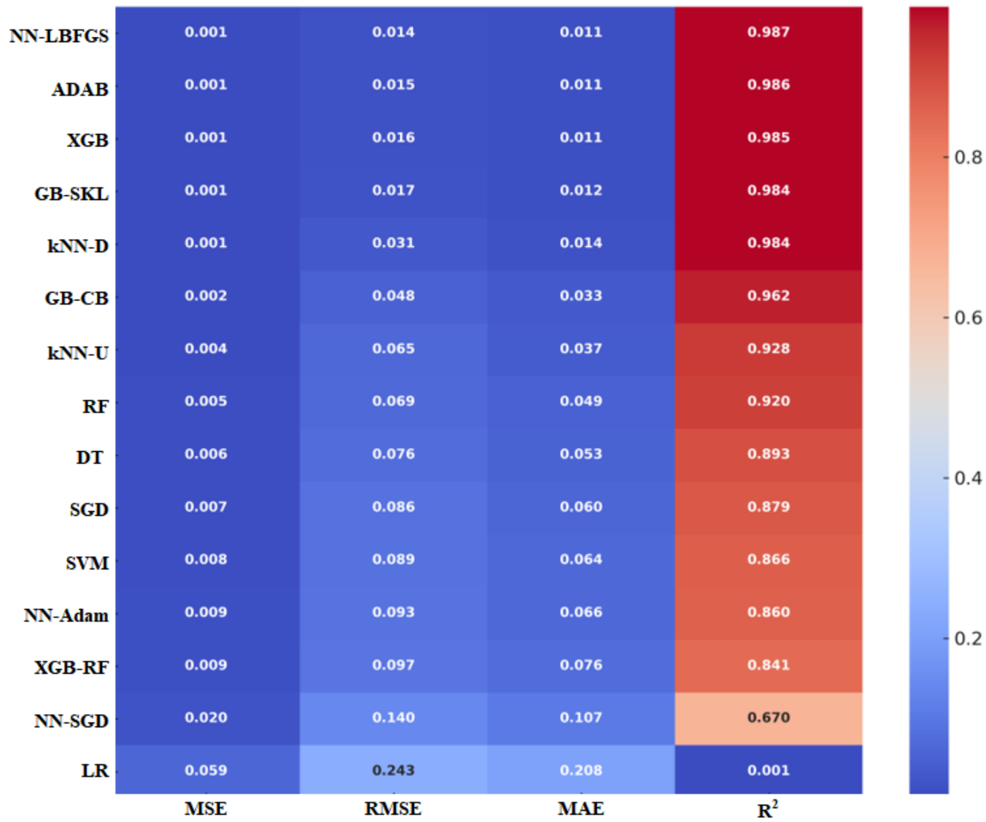

Figure 22. Performance heatmap for sixteen models: Neural Network L-BFGS achieved R^2 = 0.987.

Figure 23 represents the outstanding match of normalized actual measurements at coiled tubing depth and neural network L-BFGS predictions. Close spacing of the diagonal with even color distribution ascertains that there is uniform accuracy irrespective of the magnitude of pressure or the operating regime. The small scatter indicates lower bias and low variance, which is vital to effective real-time monitoring when carrying out nitrogen lifting operations.

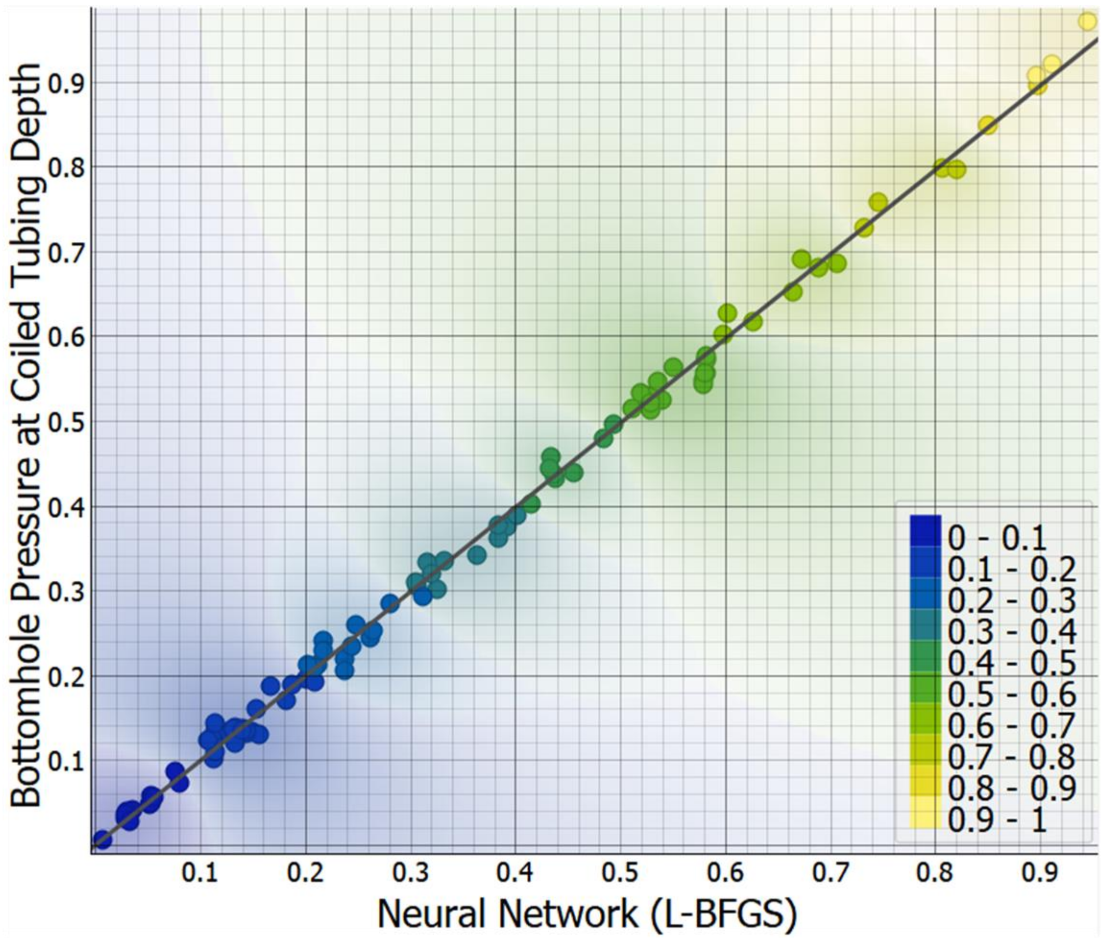

Figure 23. Normalized actual versus predicted BHP-CTD showing remarkable accuracy (R^2 = 0.987).

4.1.3.2 SHAP Analysis and Symbolic Regression

Figure 24 indicates that fluid flow rate at the surface and depth of coiled tubing are significant predictors. The SHAP range was largest in fluid flow rate, with large flow rates producing a significantly positive impact on BHP. There was a high influence of the coiled tubing depth with significant positive SHAP contributions as depth increased, which is supported by hydrostatic principles. Secondary parameters showed context-dependent contributions that were different in different operating conditions.

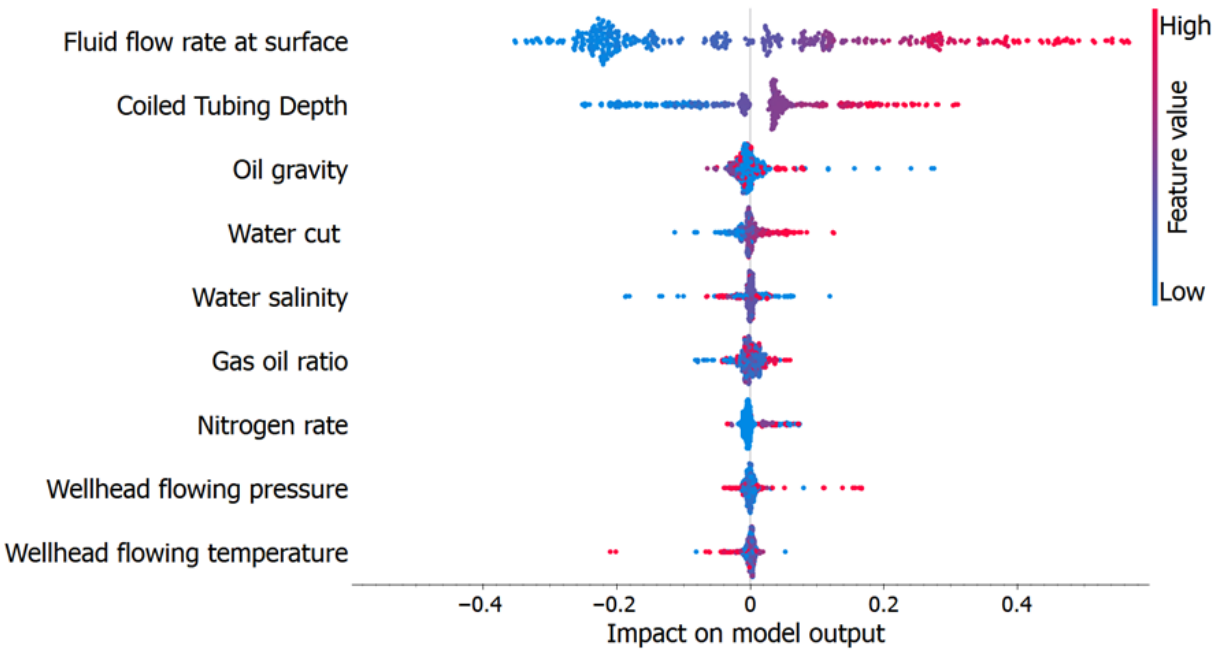

Figure 24. SHAP analysis revealing fluid flow rate and coiled tubing depth as dominant predictors.

Table 9, which represents the symbolic regression portfolio, starts with the simple expressions (complexity 1) to the complex expressions (complexity 20). Equation (8) was the best equation, which showed $R^2 = 0.94$, MSE = 0.004, RMSE = 0.063, and MAE = 0.038 at a complexity of 17 and loss = 0.00303. The formula includes fluid flow rate, coiled tubing depth, water cut, gas-oil ratio, and water salinity by transforming them, which offers an interpretable alternative, allowing their use in field control systems without machine learning infrastructure without losing competitiveness.

The variables in the symbolic regression equations below are in dimensionless form and have been normalized to the range [0, 1] using the min-max normalization constants in Section 3.4.2. Inputs to these equations need to be normalized and the output denormalized to compute bottomhole pressure in engineering units.

$$\sqrt{\cos(\mathrm{WS}) \cdot \Big(\mathrm{CTD} \cdot \big((((\mathrm{WC} + \mathrm{GOR}) + \mathrm{CTD}) \cdot \mathrm{WS}) + \mathrm{FFR\_S}) \cdot \mathrm{FFR\_S}\big)\Big)} \tag{8}$$

Table 9. Portfolio of symbolic regression equations spanning complexity 1–20.

Complexity	Loss	Equation
1	0.02577	FFR_S
2	0.0204	sin (FFR_S)
3	0.01902	sin (sin (FFR_S))
4	0.00592	$\sqrt{\mathrm{CTD} \cdot \mathrm{FFR\_}S}$
6	0.00484	$(\mathrm{FFR\_S} + 0.1697795) \cdot \sqrt{\mathrm{CTD}}$
7	0.00447	(CTD + 0.26175192) · (FFR_S + 0.12106326)
8	0.00429	sin ((FFR_S + 0.112157054) · (CTD + 0.32676274))
10	0.00396	$\sqrt{\mathrm{FFR\_S} \cdot (((\mathrm{FFR\_S} \cdot \mathrm{GOR}) + 0.85457504) \cdot \mathrm{CTD})}$
11	0.00364	$\sqrt{\left(\mathrm{FFR\_S} \cdot \left(\left(\mathrm{FFR\_S} \cdot \sqrt{\mathrm{GOR}}\right) + 0.75525254\right)\right) \cdot \mathrm{CTD}}$
13	0.00353	$\sqrt{\mathrm{CTD} \cdot (\mathrm{FFR\_S} \cdot ((\mathrm{FFR\_S} \cdot \sqrt{\mathrm{GOR}}) + \sin(\cos(\mathrm{WHT}))))}$
14	0.00346	$\sqrt{(\mathrm{CTD} \cdot \mathrm{FFR\_S}) \cdot \left(\left(\mathrm{FFR\_S} \cdot \sqrt{\mathrm{GOR}}\right) + \cos(\sin(\sqrt{\mathrm{OG}}))\right.}$
15	0.0034	$\sqrt{((\mathrm{FFR\_S} \cdot \sqrt{\mathrm{GOR}}) + \cos(\sqrt{\mathrm{OG}})) \cdot ((\mathrm{FFR\_S} \cdot \mathrm{CTD}) + 0.0032517365)}$
16	0.00335	$\sqrt{(\cos(\sqrt{\sin(\mathrm{OG})}) + (\mathrm{FFR\_S} \cdot \sqrt{\mathrm{GOR}})) \cdot ((\mathrm{CTD} \cdot \mathrm{FFR\_S}) + 0.0026609995)}$
17	0.00303	$\sqrt{\cos(\mathrm{WS}) \cdot \left(\mathrm{CTD} \cdot \left((((\mathrm{WC} + \mathrm{GOR}) + \mathrm{CTD}) \cdot \mathrm{WS}) + \mathrm{FFR\_S}) \cdot \mathrm{FFR\_S}\right)\right)}$
18	0.00296	$\sqrt{\cos(\mathrm{WS}) \cdot \left(\mathrm{CTD} \cdot \left((((\mathrm{WC} + \sin(\mathrm{GOR})) + \mathrm{CTD}) \cdot \mathrm{WS}) + \mathrm{FFR\_S}) \cdot \mathrm{FFR\_S}\right)\right)}$
19	0.00284	$\sqrt{\cos(\mathrm{WS}) \cdot \left(\left((\mathrm{FFR\_S} + ((\mathrm{CTD} + (\mathrm{WC} + \mathrm{GOR})) \cdot \mathrm{WS})) \cdot (\mathrm{FFR\_S} \cdot \mathrm{CTD})\right) + 0.0034129177\right)}$
20	0.00271	$\sqrt{\cos(\mathrm{WS}) \cdot \left(\left((\mathrm{FFR\_S} + (((\mathrm{CTD} + \mathrm{WC}) + \mathrm{GOR}) \cdot \sin(\mathrm{WS}))) \cdot (\mathrm{FFR\_S} \cdot \mathrm{CTD})\right) + 0.0048273\right.}$

4.2 Model Testing and Validation

The quality and the scalability of the machine learning models created to work on all three applications were carefully tested with systematic validation processes. Two cross-validation methods were used: k-fold cross-validation and repeated random sampling cross-validation (RRSCV). The methodologies offer detailed evaluation of model performance beyond mere train-test splits to identify any overfitting behaviors and also to provide predictive accuracy on unknown data.

4.2.1 Hydraulic Fracturing Application

4.2.1.1 K-Fold Cross-Validation Results

In K-fold cross-validation, the dataset is divided into K subsets (folds), with K-1 folds being used to train and the remaining one fold being used to test. This is repeated K times, where each fold is used as the test set. In case of the hydraulic fracturing application, 10-fold cross-validation was applied on the 42-well data set. The means of the performance measures of all folds are in Table 10 and indicate that the gradient boosting model retained its high performance with the above R^2 = 0.912, MSE = 0.004, RMSE = 0.06, and MAE = 0.041. Both the decision tree and neural network (L-BFGS) models also had high levels of cross-validation expressed in R^2 of 0.894. This method of systematic validation ensured that the predictive accuracy of the models was not due to the accidental distribution of data but that it was as a result of real learning of underlying physical relationships.

Table 10. K-fold cross-validation performance metrics for machine learning models predicting bottomhole pressure during hydraulic fracturing operations (K=10, dataset: 42 wells, 485,442 data points).

Model	MSE	RMSE	MAE	R^2

Gradient boosting	0.004	0.06	0.041	0.912
Neural network (L-BFGS)	0.004	0.063	0.045	0.894
Tree	0.004	0.063	0.042	0.894
Neural network (Adam)	0.004	0.064	0.047	0.889
kNN	0.004	0.064	0.042	0.889
Random forest	0.004	0.066	0.043	0.883
Neural network (SGD)	0.005	0.07	0.053	0.868
AdaBoost	0.005	0.072	0.046	0.86
Linear regression	0.006	0.078	0.06	0.834
SGD	0.007	0.082	0.063	0.818
SVM	0.02	0.14	0.103	0.471

4.2.1.2 Repeated Random Sampling Cross-Validation

The RRSCV procedure was performed by randomly dividing the dataset into training (80%) and testing (20%) datasets over 10 independent runs. RRSCV also adds stochastic variation to the systematic partitioning found in k-fold validation, which is more realistic in the situation of actual deployment. The findings of Table 11 supplement the k-fold results, and gradient boosting attains $R^2 = 0.917$, MSE = 0.003, RMSE = 0.059, and MAE = 0.041. The agreement between k-fold and RRSCV measurement offers good links to the stability of the model in various strategies of sampling data.

Table 11. Repeated random sampling cross-validation results for hydraulic fracturing BHP prediction models (10 iterations, 80/20 train-test split).

Model	MSE	RMSE	MAE	R^2
Gradient Boosting	0.003	0.059	0.041	0.917
Tree	0.004	0.062	0.041	0.897
Neural Network (L-BFGS)	0.004	0.063	0.045	0.894
kNN	0.004	0.064	0.042	0.891
Neural Network (Adam)	0.004	0.065	0.048	0.887
Random Forest	0.004	0.065	0.043	0.885
Neural Network (SGD)	0.004	0.067	0.05	0.88
AdaBoost	0.005	0.069	0.044	0.874
Linear Regression	0.006	0.078	0.059	0.836
SGD	0.007	0.081	0.061	0.822
SVM	0.026	0.162	0.131	0.302

4.2.1.3 Model Stability and Robustness Assessment

The fact that the performance measures are relatively similar in all validation folds means that there is very strong model stability. The gradient boosting, decision tree, and neural network (L-BFGS) models were especially robust, with R^2 values of 0.90 and above holding across all the cases of validation. This stability is essential in field deployment where there might be some deviation between the operational conditions and the training data distributions.

4.2.2 Gas Lift Wells Application

4.2.2.1 K-Fold Cross-Validation Performance

In the case of the gas lift application, where 304 wells were used, 10-fold cross-validation showed that the neural network with the L-BFGS optimizer was able to retain high performance (Figure 25). Cross-validation of the model determined R 2 = 0.960, MSE = 0.002, RMSE = 0.04, and MAE = 0.026, which reflected only slight deterioration of the original test set results (R 2 = 0.965). ADAB (R^2 = 0.905) and XGB (R^2 = 0.909) were also found to be competitive models that had the ability to capture the complex multiphase flow characteristic in gas lift systems.

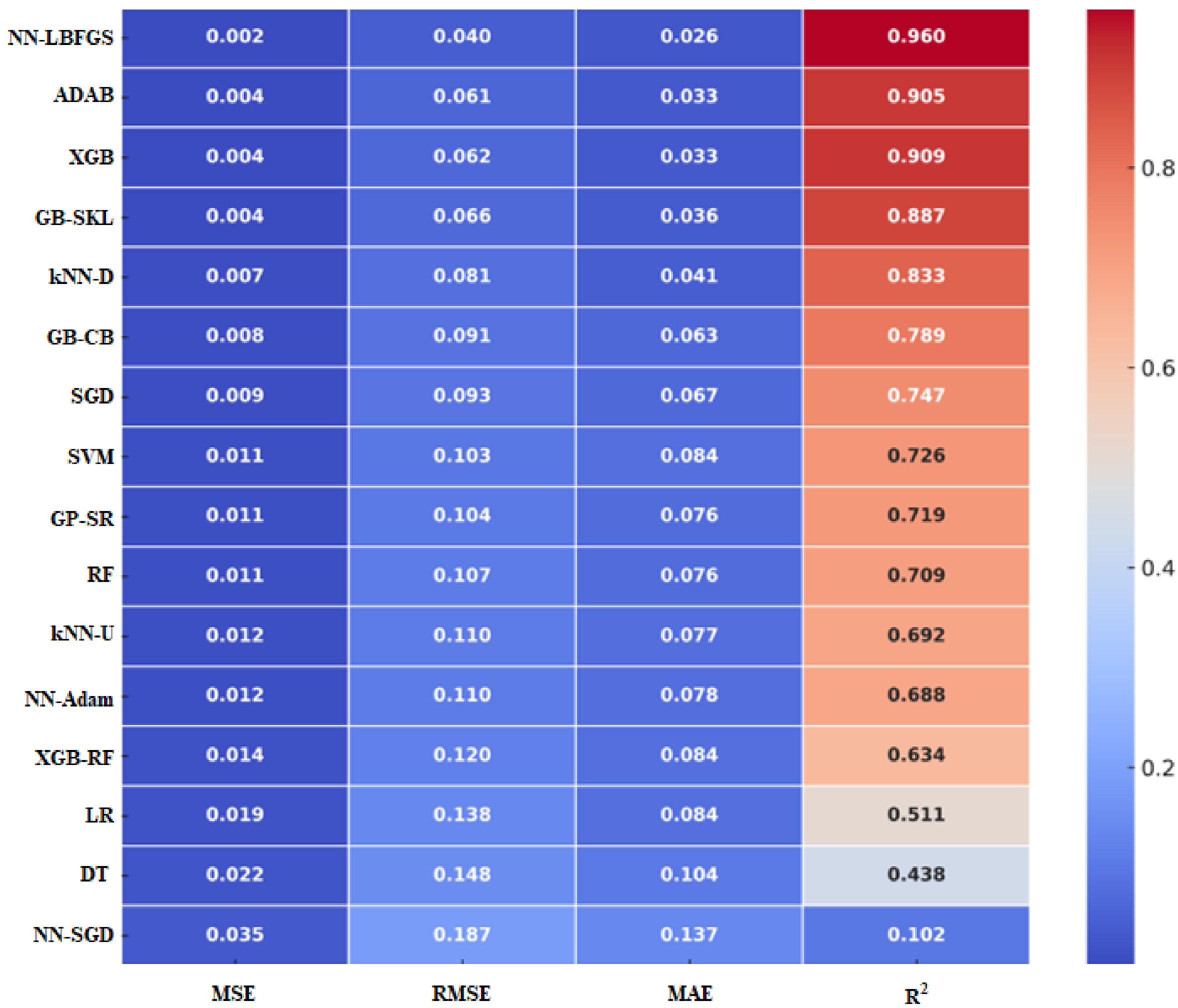

Figure 25. K-fold cross-validation performance evaluation of machine learning models for gas lift well bottomhole pressure prediction across ten folds.

4.2.2.2 Repeated Random Sampling Validation

The results of cross-validation were supported with the RRSCV procedure of 10 iterations (Figure 26). The neural network (L-BFGS) returned the following results: R^2 = 0.931, MSE = 0.003, RMSE = 0.053, and MAE = 0.032. The fact that k-fold and RRSCV are closely similar across several models demonstrates the fact that the performance does not depend on the data partitioning scheme, hence increasing confidence in deployment reliability.

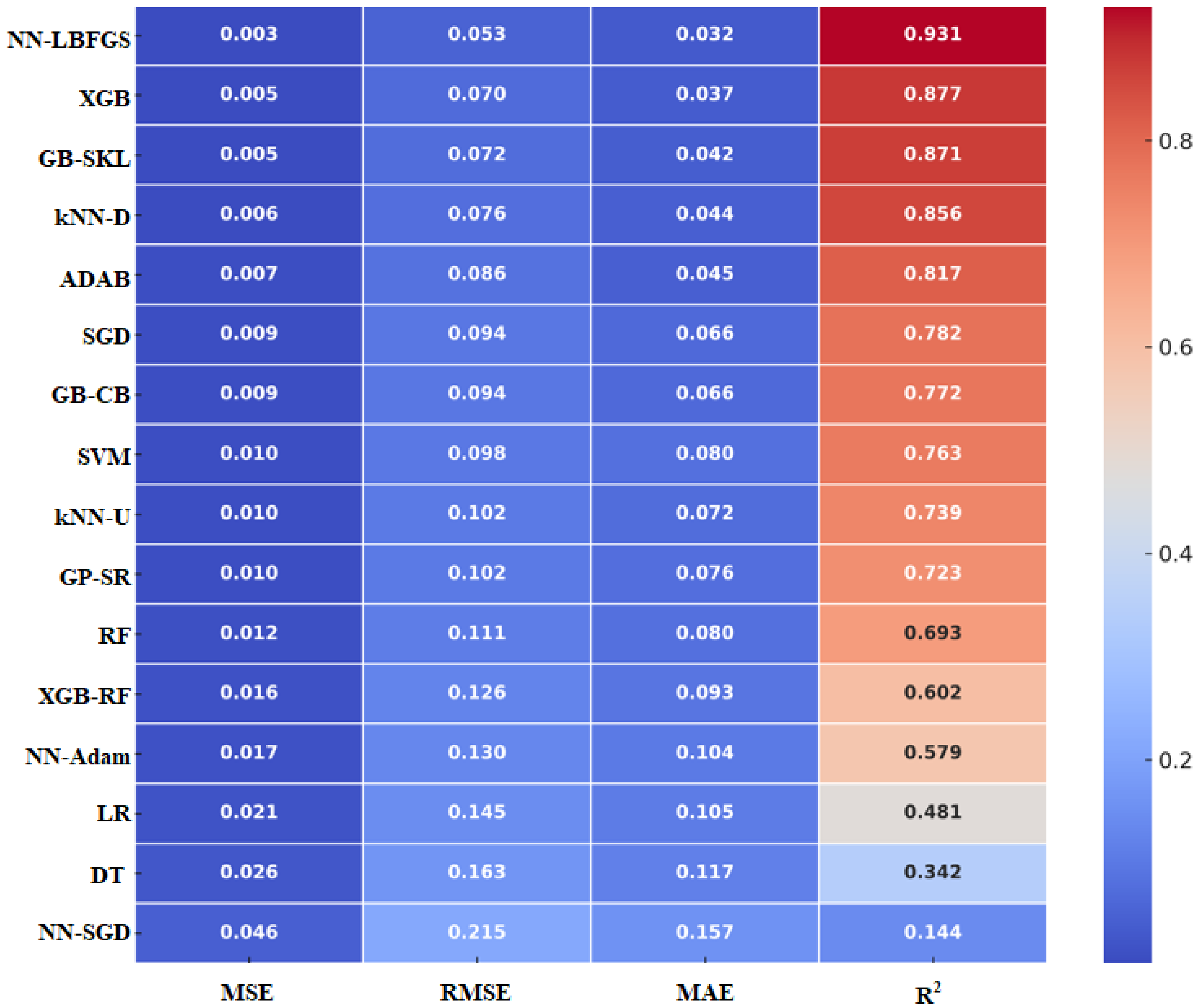

Figure 26. Performance metrics from repeated random sampling cross-validation for gas lift BHP prediction models (10 iterations).

4.2.2.3 Model Stability and Robustness Assessment

The gas lift models were shown to be very consistent among validation methods. This stability is of special interest considering the operational diversity that is included in the dataset, such as differences in the properties of the reservoirs, the configuration of the completion, and gas injection parameters.

4.2.3 Nitrogen-Lifted Wells Application

4.2.3.1 K-Fold Cross-Validation Results

The nitrogen-lifted wells dataset that included 518 wells was subjected to rigorous 10-fold cross-validation (Figure 27). The neural network using the L-BFGS optimizer had $R^2 = 0.983$, MSE = 0.030, RMSE = 0.064, and MAE = 0.016, indicating that it performed equally across folds. Extreme gradient boosting (GB-SKL) and XGB models also demonstrated good cross-validation performances with the values of R^2 being 0.974, which indicates that ensemble models can effectively deal with the multiphase flow effects in the nitrogen lifting processes.

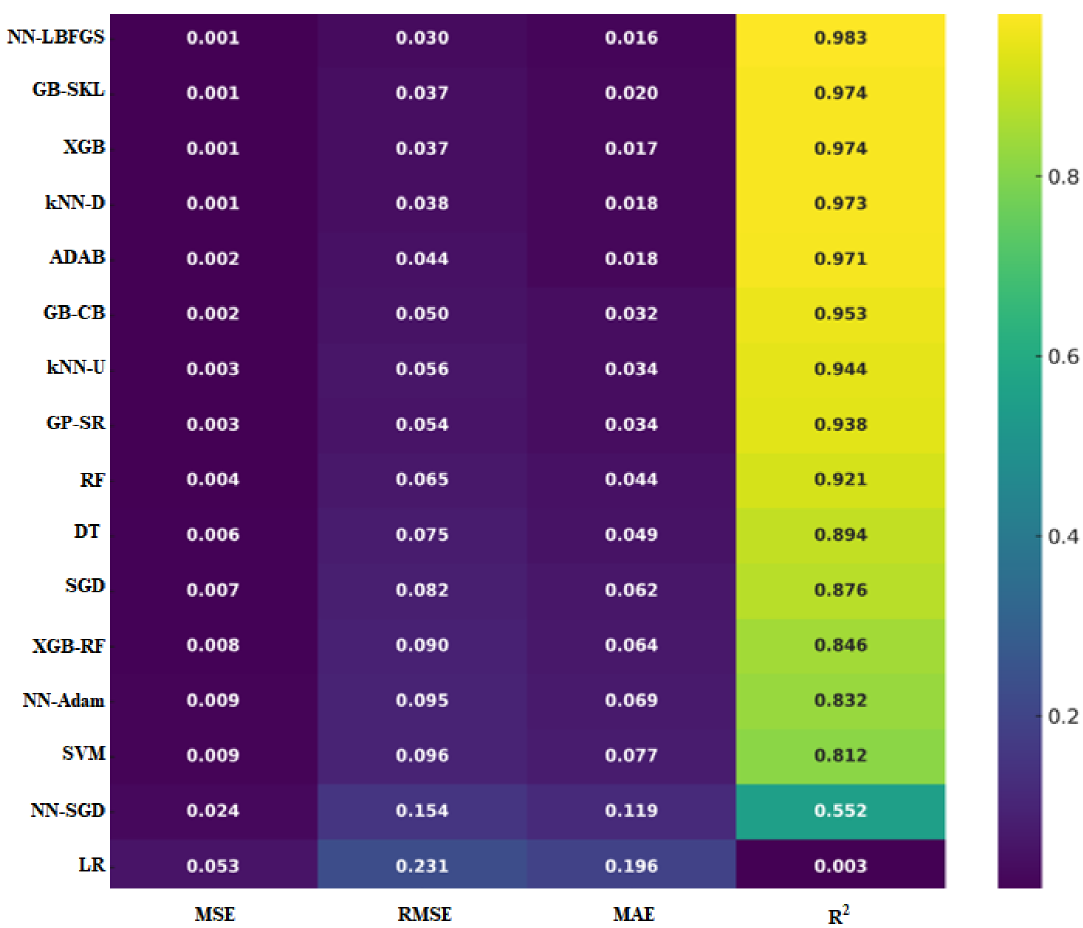

Figure 27. K-fold cross-validation results demonstrating model performance stability for nitrogen-lifted wells BHP prediction (K=10).

4.2.3.2 Repeated Random Sampling Performance

The output of RRSCV with 10 iterations is shown in Figure 28, where the neural network (L-BFGS) attains the following values: $R^2 = 0.983$, MSE = 0.001, RMSE = 0.031, and MAE = 0.016. The symbolic regression model too showed consistency in performance with $R^2 = 0.931$ at validation repetitions and offers a model that is interpretable with an acceptable accuracy loss of about 5 percent compared to the best-performing black-box model.

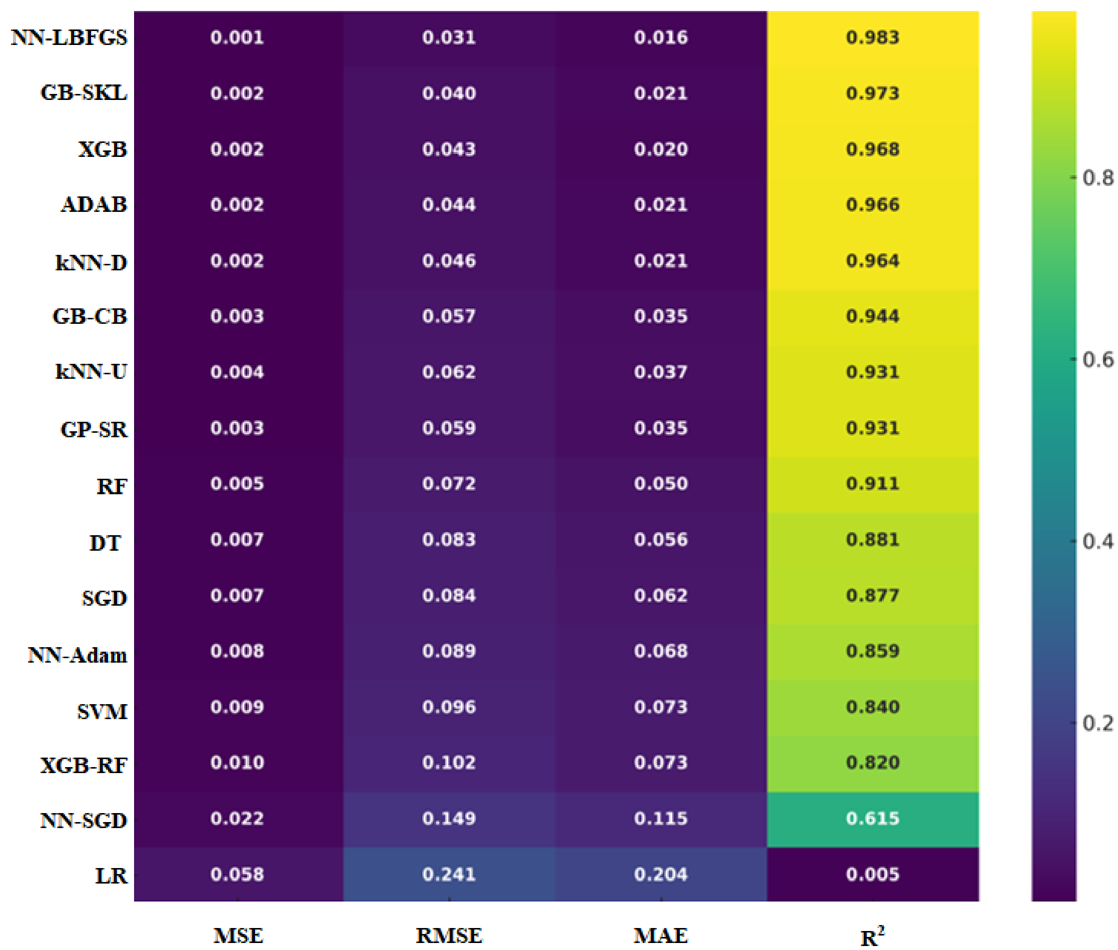

Figure 28. Repeated random sampling validation outcomes for nitrogen-lifted wells bottomhole pressure prediction models across ten independent iterations.

4.2.3.3 Model Stability and Robustness Assessment

The models of the nitrogen-lifted wells had very high levels of robustness, and their performance measurements had very small differences when using different validation methodologies. The agreement between k-fold and RRSCV data, together with a successful blind test on 29 independent wells, is a good indication that the models have learned generalizable patterns and not artifacts of 29 independent wells. The constant performance of the neural network (L-BFGS) model in various validation conditions makes it a confident tool in predicting bottomhole pressure in real-time during nitrogen lifting operations.

4.3 Field Application

The final test of predictive models is their role in actual conditions of operations. This section reports the field application outcomes of the three domains of bottomhole pressure prediction, showing that the developed machine learning models can be practically used and deployed in real scenarios in terms of independent well testing, blind dataset verification, and comparative benchmarking with traditional models.

4.3.1 Hydraulic Fracturing Application

4.3.1.1 Well X Case Study

A gradient boosting model with the best performance was applied to Well X, a vertical well drilled in the Western Desert of Egypt, 8,500 feet in a conventional sandstone reservoir (Formation Y), to test its applicability in practice. Table 12 outlines the operational and completion parameters of the well, such as the tubing inner diameter (2.992 inches), depth of pressure gauge (7,950 feet), gel load (40 lb/mgal), proppant size (20/40 mesh, specific gravity 3.544), and other surface parameters of operation that were measured during the hydraulic fracturing treatment.

Table 12. Well-X input parameters.

Parameter	Units	Value
Tubing inside diameter	inches	2.992
Pressure gauge depth	ft	7950
Gel load	lb/mgal	40
Proppant size	mesh	20/40
Proppant speicific gravity	-	3.544

4.3.1.2 Real-Time Prediction Accuracy

The gradient boosting model was trained to handle real-time streams of data throughout the fracturing process to use available surface pressure, slurry flow rate, surface proppant concentration, and other conveniently available parameters to produce continuous predictions of bottomhole pressure. The model proved to be very accurate in prediction during the treatment stages, starting with initial pad pumping up and proppant-laden slurry injection to the end of the treatment.

4.3.1.3 Comparison with Downhole Gauge Measurements

The overall results of the comparison between the gradient boosting model predictions and the actual bottomhole pressure measurements of the installed memory gauge during the entire process of fracturing Well X are shown in Figure 29. As can be seen in the visualization, an outstanding consistency between the predicted and measured values is observed among all ranges of treatment and also during transitions between clean fluid and proppant stages where pressure dynamics are the most complicated. The model was effective in tracing the changes in the pressure caused by the alteration of pumping rate, proppant concentration, and fluid rheology, enabling it to demonstrate its ability to trace the nonlinear forms of relationships between the bottomhole pressure and the operation of hydraulic fracturing.

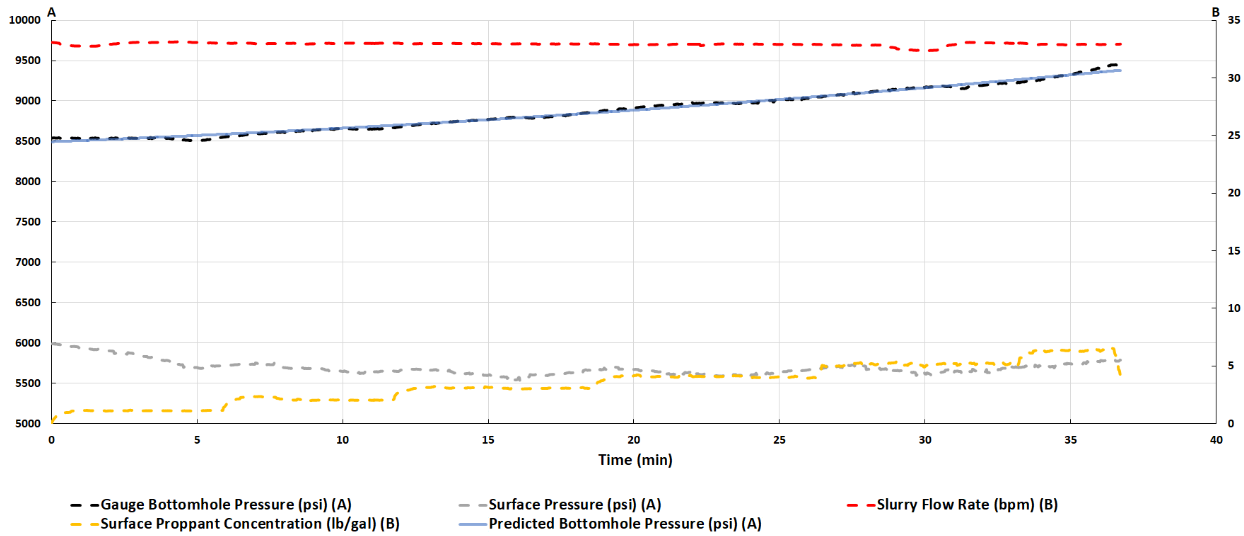

Figure 29. Real-time monitoring of hydraulic fracturing treatment in well A.

4.3.1.4 Operational Insights and Optimization Opportunities

The field use on Well X has shown that the gradient boosting model can become a potential tool in real-time monitoring and make it possible to perform continuous treatment performance assessment of hydraulic fracturing without using the expensive and operationally difficult downhole pressure gauges. This allows quick optimization actions during pumping events, such as a change in gel injection rates, proppant concentration ramping schedules, and identifying the possibility of a screenout early. The capability of the model to give real-time bottomhole pressure estimates allows the engineers to optimize the fracture geometry in real time, which could enhance the effectiveness of the treatment and also save on operation costs and risks that would arise when the treatment is terminated or prolonged past the optimal operating environment.

4.3.2 Gas Lift Well Applications

4.3.2.1 Blind Dataset Validation

The most rigorous test of model generalizability involves evaluation on completely independent data never encountered during training, hyperparameter tuning, or feature selection processes. To this end, the neural network (L-BFGS) model was validated on a blind dataset comprising 30 gas lift wells with distinctly different operational characteristics and reservoir properties from the training set. Table 13 presents comprehensive descriptive statistics for this validation dataset, spanning fluid flow rates from 50 to 1,794 STB/D, water cuts from 0% to 99%, gas-oil ratios from 100 to 1,500 SCF/STB, wellhead flowing pressures from 30 to 250 psi, and injection point depths from 2,697 to 9,253 feet. This diversity in operational parameters provides a stringent test of model robustness across the operational envelope encountered in field applications.

Table 13. Descriptive statistics for the dataset from 30 gas lift wells.

Parameter	Units	MIN	MAX	AVG	Median
Tubing inside diameter	inches	2.441	2.992	2.7165	2.7165
Tubing depth	ft	3200	10,318	6598.633	6220
Perforation depth	ft	3250	10,368	6686.3	6343
Wellhead pressure	psi	30	250	114.3333	100
Reservoir temperature	f	112.5	183.68	146.863	143.43
Gas–oil ratio	scf/stb	100	1500	582.5	500
Water cut	%	0	99	75.66667	90
Operating gas injection pressure	psi	800	1100	905	900
Gas lift injection rate	mmscf/d	0.4	0.9	0.52	0.5
Bottomhole pressure	psi	448	1661	800.2	745.5
Injection point depth	ft	2697	9253	5737.433	4799
Fluid flow rate	stb/d	50	1794	780.3667	734.5

4.3.2.2 Generalization Performance Assessment

Figure 30 shows that the blind dataset validation has an impressive ability to generalize, as the neural network (L-BFGS) has a high $R^2 = 0.97$ when predicting the bottomhole pressure of the 30 independent wells. The fact that this performance was obtained without any retraining of the models or changing of the parameters confirms that the learned representations are able to provide the basic physical relations that govern multiphase flow in gas lift systems and are not an artifact of the data. Assistive consistency in test set performance ($R^2 = 0.965$), cross-validation performance ($R^2 = 0.960$), and blind testing performance ($R^2 = 0.97$) is a good indicator of actual model generalizability.

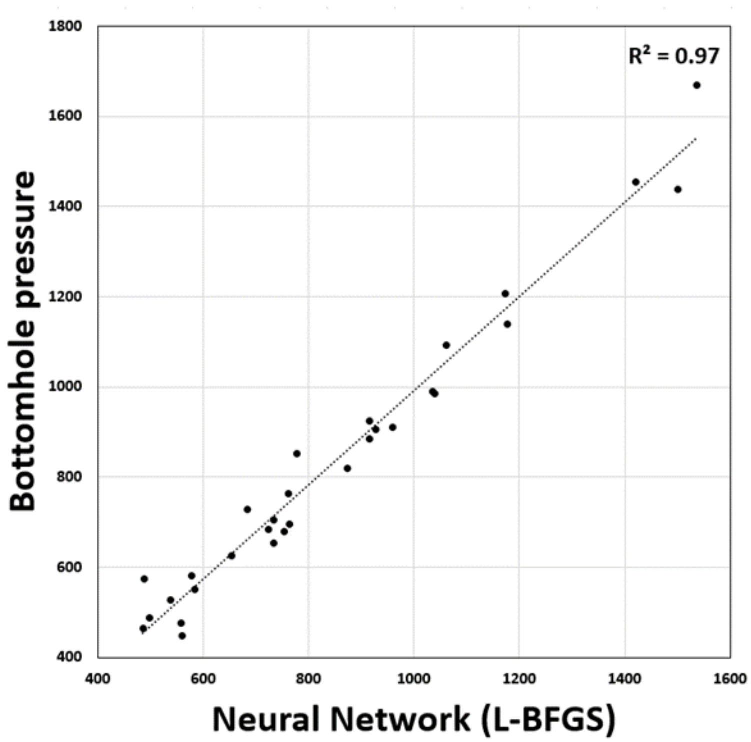

Figure 30. Neural network (L-BFGS optimizer) prediction vs. actual BHP values across 30 gas lift wells.

4.3.2.3 Field Implementation Readiness

The effective validation of the blind dataset proves the fact that the neural network (L-BFGS) model is prepared to be implemented in the field during gas lift operations. All input parameters needed, including tubing inside diameter, tubing depth, perforation depth, wellhead pressure, reservoir temperature, gas-oil ratio, water cut, operating gas injection pressure, gas lift injection rate, injection point depth, and fluid flow rate, are all periodically measured and available in real-time using standard field SCADA systems. The computational effectiveness of the model allows it to easily fit into the current operational processes with no extra hardware and no delays in the decision-making process.

4.3.2.4 Real-Time Monitoring Capabilities

The accuracy and computational efficiency of the neural network (L-BFGS) model place it as a powerful aid in the continuous monitoring of bottomhole pressure in gas lift activities. With this ability, operators are able to optimize the rate of gas injection dynamically, identify operational anomalies, and make decisions on the changes to the artificial lift system or well stimulation needs. The removal of the need to use costly and hard-to-install downhole pressure gauges is an important operating and economic benefit, especially in mature fields where many gas lift wells are present and need regular performance evaluation.

4.3.3 Nitrogen-Lifted Wells Application

4.3.3.1 Blind Testing Results

The neural network (L-BFGS) model underwent rigorous blind testing on 29 vertical oil wells completely independent from the 518-well training dataset. Table 14 documents the substantial operational diversity in this validation set, with bottomhole pressures ranging from 851 to 3,783 psi, fluid flow rates from 88 to 3,573 STB/D, water cuts from 0% to 100%, coiled tubing depths from 3,000 to 13,028 feet, and nitrogen rates from 400 to 750 SCF/min. This comprehensive parameter range ensures robust evaluation of model performance across realistic field conditions encountered in nitrogen lifting operations.

Table 14. Dataset statistics from a sample of 29 wells

Parameter	Units	MIN	MAX	AVG	Median
Bottomhole pressure at coiled tubing depth	PSI	851	3783	2304	2522
Fluid flow rate at surface	STB/D	88	3573	1437	1241
Water cut	%	0	100	37	30

Gas–oil ratio	SCF/STB	0	1556	611	500
Water salinity	PPM	51,000	200,000	143,451	150,000
Wellhead flowing pressure	PSI	30	490	97	67
Wellhead flowing temperature	F	90	117	104	107
Coiled tubing depth	FT	3000	13,028	8628	9002
Nitrogen rate	SCF/M	400	750	584	600
Oil gravity	API	22	46	38	35

4.3.3.2 Benchmarking Against Traditional Correlations

One of the essential strengths of the blind testing methodology that has been used in this research is that it can test the neural network model and the traditional empirical correlations at the same time and under the same conditions. The five empirical correlations were computed using PROSPER software. Figure 31 is a detailed six-panel prediction versus actual bottomhole pressure at depth of coiled tubing of (a) neural network (L-BFGS), (b) Hagedorn-Brown correlation, (c) Beggs-Brill correlation, (d) Orkiszewski correlation, (e) Fancher-Brown correlation, and (f) Duns-Ros correlation. The visual comparison shows the high quality of the machine learning method at a glance as the data values cluster tightly around the diagonal line in the case of the neural network prediction, whereas the traditional correlations are widely scattered with a significant systematic variance between the measured values and the traditional correlations.

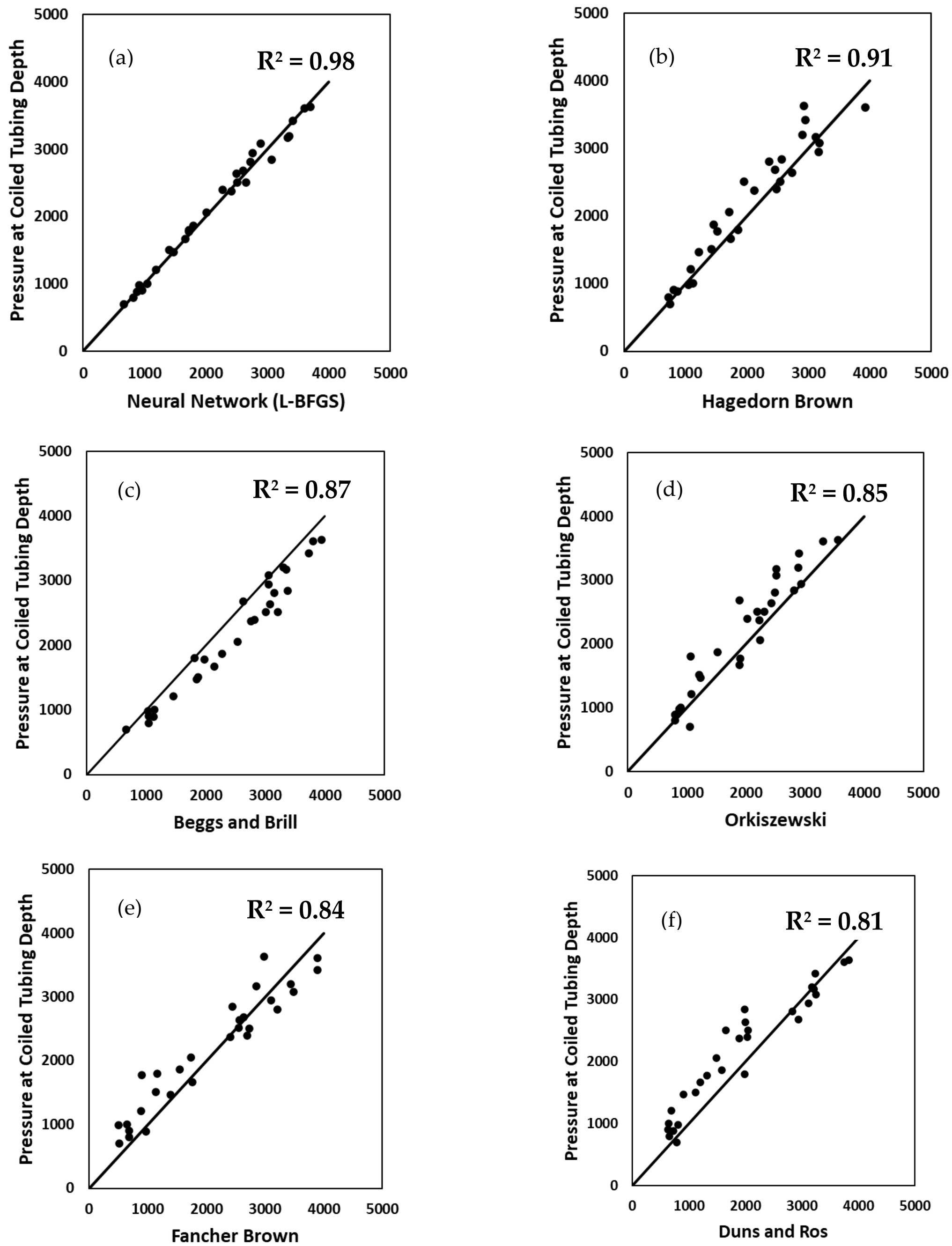

Figure 31. Comparison of measured pressures at coiled tubing depth with predicted pressures from various methods: (a) NN-LBFGS, (b) Hagedorn–Brown, (c) Beggs–Brill, (d) Orkiszewski, (e) Fancher–Brown, and (f) Duns–Ros.

4.3.3.3 Quantitative Superiority Analysis

Table 15 gives quantitative performance measures of each of the six prediction methods considered on the 29-well blind dataset. The neural network (L-BFGS) was able to generate MSE = 9,791, RMSE = 99 psi, MAE = 76 psi, and $R^2 = 0.98$, which is very accurate in terms of absolute pressure units. By contrast, the highest performing traditional correlation (Hagedorn-Brown) had MSE = 74,600, RMSE = 273 psi, MAE = 212 psi, and $R^2 = 0.91$, which is an 86% and 179% improvement in mean squared error and mean absolute error, respectively, over the neural network. The worst correlation (Duns-Ros) had MSE = 155,331, RMSE = 394 psi, MAE = 325 psi, and $R^2 = 0.81$ and contained more than fifteen times the errors of the machine learning model. This quantitative study conclusively shows that data-based methods are better than the conventional empirical techniques in the prediction of bottomhole pressure in nitrogen-lifted wells.

Table 15. Evaluation metrics for BHP-CTD prediction methods using MSE, RMSE, MAE, and R2.

BHP-CTD Prediction Methods	MSE	RMSE	MAE	R^2
Neural Network (L-BFGS)	9791	99	76	0.98
Hagedorn and Brown	74,600	273	212	0.91
Beggs and Brill	107,223	327	277	0.87
Orkiszewski	117,194	342	272	0.85
Fancher and Brown	127,148	357	295	0.84
Duns and Ros	155,331	394	325	0.81

4.3.3.4 Real-Time Implementation Potential

Table 16 summarizes the comparative benefits and drawbacks of machine learning models and traditional models, which is a fair evaluation of deployment aspects. The machine learning methodology is accurate, flexible to a wide range of well and reservoir conditions, can be applied in real-time, is resistant to missing or noisy data in well-prepared forms, and has interpretable symbolic regression options. Although computationally easy and well-established in the industry, traditional empirical correlations have poor accuracy outside original calibration regimes, poor generalizability, and are incapable of modeling multidimensional interactions. The physically grounded and interpretable nature of mechanistic models is accompanied by large input parameter requirements, large calibration effort, and computational resource needs, which restrict the real-time field application capability. The fact that the neural network model has lightweight computational needs and only requires the routinely measured operational parameters makes it exceptionally qualified to be integrated into the existing SCADA systems, allowing continuous bottomhole pressure monitoring in real-time to operate without incurring any extra costs on instruments and without extra complexity in implementation.

Table 16. Comparative advantages and limitations of machine learning models versus traditional methods for BHP-CTD prediction.

Approach	Advantages	Limitations
Machine Learning Models	▪ High accuracy ▪ Flexibility with a variety of well/reservoir conditions ▪ Capability for real-time application ▪ Resistant to missing/noisy data when preprocessed ▪ Symbolic regression provides interpretable equations	▪ Reliance on massive datasets of good quality ▪ Possibility of retraining on the change in conditions ▪ There are such models (e.g., neural networks), which are a black box
Empirical Correlations	▪ Easy to apply ▪ Low computation demands ▪ Well recognized, and established in industry	▪ Poor accuracy beyond original calibration range ▪ Weak generalizability to variety of well conditions ▪ Failure to model multidimensional interactions
Mechanistic Models	▪ Physically grounded and interpretable ▪ Able to simulate multiphase flow regimes	▪ Demand many input parameters and calibration ▪ Computationally demanding (particularly OLGA)

	▪ Widely validated in academic and industrial contexts	▪ Limited feasibility for real-time field application

4.4 Limitations of Machine Learning Models for Bottomhole Pressure Prediction

Although the designed machine learning models were shown to have a high predictive accuracy in all three applications, it is important to note that they have numerous intrinsic limitations, which should be considered to deliver a balanced evaluation of their implementation opportunities and restrictions.

4.4.1 Hydraulic Fracturing Application

Hydraulic fracturing BHP prediction models have weaknesses that are mainly associated with the extent of the data set and generalizability of operations. The training dataset, which consists of 42 vertical wells in the Western Desert of Egypt, has a variety of operating and completion conditions, but the predictive performance of the models beyond these defined ranges of these parameters has not been validated. Live BHP forecasting shows that proper quality of historical data is highly valued since error in sensor measurements and fast changes in pressure during treatment steps can be reflected in the prediction pipeline, affecting its accuracy at important operation periods.

One of the key constraints of the machine learning approach is lack of interpretability as compared with physics-based models. In contrast to mechanistic formulations based on the principles of fundamental fluid mechanics, gradient boosting and neural network models are complex pattern recognition models without explicit mathematical foundations that can directly relate input parameters to physical phenomena. This black-box feature makes it more difficult to explain individual predictions to field engineers and can make it difficult to accept within the organization with strong concerns over physically transparent modeling methods.

4.4.2 Gas Lift Well Application

The gas lift well models have similar data-based limitations but have application-specific challenges. The performance of the model is highly sensitive to the quality and quantity of data, and degraded predictions are observed when the model is trained on inaccurate, insufficient, or systematically biased data. The 304-well training set has quite good coverage of common gas lift operational conditions; however, wells that run at the extremes of parameter space or in uncharacteristic configurations might not be sufficiently well-covered in the learned feature space.

Gas lift systems are associated with more complex conditions as a result of operational dynamics. Alternations in the pressure of the reservoir, the characteristics of the fluids, valve characteristics, or injection plans may shift the operational envelope beyond that of the training data distributions, and thus a periodic retraining of the model is needed to ensure successful prediction accuracy. This condition imposes continuous calculational and logistical loads on fieldwork, especially when dealing with large field operations of hundreds of gas lift wells at a time.

The performance of the models is very sensitive to the hyperparameter settings and preprocessing strategies. The choice of neural network (L-BFGS) optimizer, hidden layer, regularization, and normalization choices have a strong impact on predictive accuracy. To operate this sensitivity, the initial deployment of this sensitivity requires meticulous tuning and careful monitoring to ensure that performance deterioration is detected and recalibration is performed.

4.4.3 Nitrogen-Lifted Wells Application

The models of nitrogen-lifted wells, although being more accurate as compared to the classical correlations, have challenges in their practical implementation. The consideration of the data acquisition infrastructure, real-time capability computation, and personnel training is necessary to integrate into the current operational workflows. The proposed 12-month retraining cycle or even more frequent retraining in case of large amounts of new well data requires a permanent allocation of resources to keep the models up-to-date and to re-version them.

The architectural limitations can be seen when dealing with a highly complex nonlinear relationship or when dealing with multiphase flow regime transitions, which are not sufficiently represented in training data. The simple neural network architectures will not be able to capture interaction effects that are subtle among the parameters of operation, whereas more complex architectures can easily overfit, especially when feature space is highly dimensional. The models also exhibit sensitivity to given input parameters, especially the depth of coiled tubing and rate of nitrogen, where the absence of data will need backup mechanisms like retraining to limited-input or statistical imputations to preserve operational continuity.

The basic interpretability constraint remains present throughout the nitrogen lifting application, in which practitioners cannot recreate individual predictions of the physics underlying multiphase flow behavior. Although symbolic regression yields more interpretable equations with moderately acceptable accuracy penalties ($R^2 = 0.94$ compared to 0.987 using neural networks), their equations lack the formal theoretical basis of mechanistic multiphase flow models and are thus less likely to be accepted in applications when regulatory validation or engineering rationale is needed.

In addition to the application-specific limitations listed above, three practically relevant data quality issues need to be considered to maintain reliable model predictions during deployment in the field for all three applications. For the case of noisy sensor measurements of one or more input parameters, it is recommended to implement a real-time moving average filter or Kalman filter in the SCADA integration layer to smooth out transient measurement spikes before the data are passed to the prediction pipeline, along with sensor measurement bounds validation that automatically flags any incoming sensor readings that fall outside physically reasonable operating ranges, thereby ensuring that bad measurements do not silently flow through the prediction system. For missing input parameters (due to sensor issues or communication breakdown), a stratified approach is recommended depending on the time gap. For very short-duration gaps (seconds to minutes), propagating the last known valid value can be used as a physically plausible proxy for non-varying or slowly varying input parameters. For longer-duration gaps, the prediction should be flagged as unreliable and the operator alerted rather than allowing the model to make potentially unreliable predictions without warning, and in the case of a permanently unavailable input, a reduced-feature-set model variant should be trained using the remaining parameters as input, with the SHAP analysis results used to confirm that the missing parameter is not a major predictor, thereby justifying the reduced-feature-set predictions. In the case of extreme values beyond the min-max normalization range of the training data, the model is being asked to make predictions beyond its training bounds, which results in unpredictable and potentially physically unrealistic predictions. It is therefore crucial to incorporate input range checks that automatically reject or flag predictions if any normalized input value is outside the range of zero to one and notify the operator that the current operating conditions are outside the verified boundaries of application for the model and that the model should be retrained to include such extreme operating conditions before predictions are reliable again.

Chapter 5: Conclusions

5.1 Summary of Research Findings

5.1.1 Hydraulic Fracturing BHP Prediction

In this study, ten machine learning models were successfully created and validated to predict bottomhole pressure during hydraulic fracturing processes based on the data of 42 vertical wells in the Western Desert of Egypt. The gradient boosting algorithm was the best performer with R^2 = 0.931, MSE = 0.003, RMSE = 0.058, and MAE = 0.039. Neural network (L-BFGS) models and decision tree models also showed good results with R^2 of 0.903 and 0.901, respectively. The models employed easily obtained operational parameters such as surface pressure, slurry flow rate, surface proppant concentration, inside tube distance, pressure gauge depth, gel load, proppant size, and specific gravity to predict bottomhole pressure at perforation depth during guar cross-linked fluid injection.

5.1.2 Gas Lift Well BHP Prediction

Sixteen machine learning models, including genetic programming-based symbolic regression, were developed and rigorously evaluated using data from 304 gas lift wells. The L-BFGS neural network had the best results of R^2 = 0.965, MSE = 0.001, RMSE = 0.037, and MAE = 0.021 on the test set. The SHAP analysis indicated that the main factors that determine the bottomhole pressure in gas lift systems are the depth of the injection point, the tubing depth, and the rate of fluid flow. It is important to note that the symbolic regression produced a readable mathematical equation with R^2 = 0.80 that gives an understandable alternative between precision and interpretability. Robust generalizability was tested on 30 independent wells using blind validation with R^2 = 0.97.

5.1.3 Nitrogen-Lifted Well BHP Prediction

The application of nitrogen-lifted wells included the creation of sixteen models based on data of 518 vertical oil wells. The neural network (L-BFGS) had a better predictive ability of $R^2 = 0.987$, MSE = 0.001, RMSE = 0.014, and MAE = 0.011. A blind test on 29 independent wells produced R 2 = 0.98, which is very much higher than traditional empirical correlations, such as Hagedorn-Brown (R 2 = 0.91), Beggs-Brill (R 2 = 0.87), Orkiszewski (R 2 = 0.85), Fancher-Brown (R 2 = 0.84), and Duns-Ros (R 2 = 0.81). Symbolic regression was also found to give an interpretable equation with $R^2 = 0.94$, showing only a small decrease in accuracy over black-box models.

5.2 Key Contributions to Knowledge

5.2.1 Methodological Advancements

The study contributes to the development of petroleum engineering techniques by illustrating how models are selected systematically, how they are rigorously tested, and how extensive benchmarking of families of algorithms takes place. The introduction of k-fold cross-validation, repeated random sampling, and blind dataset testing creates a solid framework for the assessment of machine learning models in the subsurface. The combination of SHAP analyses offers more insight into the importance of features and model decision-making steps than at any other time, allowing the gap between predictive performance and understanding of operations to be filled.

5.2.2 Novel Applications of ML in Petroleum Engineering

This research is the first detailed implementation of sophisticated machine learning algorithms to predict the bottomhole pressure in three main key operations of petroleum engineering: hydraulic fracturing and gas lift, as well as nitrogen lifting. In contrast to the vast amount of research done in the past on single algorithms or synthetic data, this study offers a comparative analysis of 10-

16 models per app with large field datasets, which validates data-driven models as viable substitutes to physics-based models, empirical correlations, and costly downhole gauges.

5.2.3 Interpretable AI Solutions

Effective implementation of the genetic programming approach to symbolic regression overcomes the enduring criticism of machine learning as opaque black-box technology. This study shows that interpretability and performance can be matched by creating explicit mathematical models with acceptable penalties in terms of performance (loss of 10-20% R^2 versus optimal neural networks). The obtained equations give clear and non-obscure physics-equivalent equations that can be approved by regulators and justified by engineers.

5.3 Achievement of Research Objectives

5.3.1 Model Development Success

All the main research goals were achieved. In hydraulic fracturing, the bottomhole pressure can be accurately predicted based on surface-measurable parameters, and no downhole gauges are required. Gas lift models prove to have the potential to optimize injection strategies by using real-time pressure monitoring. The models of nitrogen-lifted wells are more accurate than the industry-standard correlations with an appropriate level of computational efficiency to be used in the field.

5.3.2 Validation and Testing Accomplishments

The soundness and applicability of the models were proved through rigorous validation. K-fold cross-validation showed that performance was stable with partitions of the data; repeated random sampling showed that performance was independent of stochastic variation; and blind dataset testing on 29-30 independent wells per application provided true predictive ability on untested data. The models were very accurate in a wide range of operational conditions, well configurations, and reservoir properties.

5.3.3 Field Application Demonstrations

Practical utility was proven by successful field deployments. The hydraulic fracturing model was an effective model to monitor bottomhole pressure during the Well X treatment phases, which allowed real-time optimization decisions. Nitrogen lift and gas lift models were found to be ready for continuous monitoring.

5.4 Practical Implications for Industry

5.4.1 Operational Optimization

The models developed allow real-time continuous monitoring, which facilitates dynamic optimization of treatment parameters. In hydraulic fracturing, the engineers are able to accommodate gel injection rates, proppant concentrations, and pumping schedules in response to immediate feedback of bottomhole pressure. The operations of gas lift have the advantage of an optimized injection procedure that helps in avoiding the cases of liquid loading and valve malfunctions. Nitrogen lifting operations are able to optimally control the injection rates and the positioning of coiled tubing to achieve maximum production without proppant flowback or formation damage.

5.4.2 Cost Reduction Opportunities

Machine learning models do not require the usage of costly downhole pressure gauges (average cost is $50,000-$150,000 per deployment) and related wireline or coiled tubing conveying expenses. The models only need surface-measured parameters that are readily obtained using standard field instrumentation, which are cost savings of an order of magnitude for pressure monitoring programs of large populations of wells.

5.4.3 Real-Time Decision Support

The efficiency in computation of deployed models allows real-time decision support and not post-treatment analysis. Field engineers have real-time pressure estimates during operations, which means that it is easy to respond to the changing conditions quickly, identify operational anomalies early, and avoid expensive failures, including premature screenouts or uncontrolled flowback.

5.5 Limitations of the Study

5.5.1 Data Constraints

The development of the models was based on field data in the Western Desert of Egypt, which may not be applicable to highly dissimilar geological systems, fluid systems, or working practices. Even though training datasets are large (42-518 wells per application), they might not cover all the extremes in operation or unusual forms in global field operations.

5.5.2 Model Generalizability Issues

It is not clear how performance would be outside training data parameters. Unconventional completion design, extreme conditions of operation, novel fluid systems, or wells that may use unusual fluid systems may need retraining or validation of their models prior to use. The reported ability to generalize to blind data in similar operation regimes does not presuppose its behavior in clearly different settings.

5.5.3 Implementation Challenges

Field deployment necessitates the incorporation of the preexisting SCADA systems, real-time data-acquiring infrastructure, and training of the personnel. Organizations that use outdated systems or have insufficient digital infrastructure can have technical implementation obstacles. The black-box characteristics of the most successful neural networks can be met with opposition in the regulatory space that needs transparent and physically intuitive models.

5.5.4 Computational Considerations

Although prediction inference is computationally easy, preliminary training of models demands considerable amounts of computers and expertise. Regular retraining requires continuous allocation of resources, creating versions, and ensuring quality, which can be difficult with smaller operators that do not have specific data science resources.

5.6 Recommendations for Future Research

5.6.1 Model Enhancement Opportunities

The next step in future research could include the introduction of physics-informed neural networks that directly use governing equations as constraints during training, which may allow the extrapolation of the prediction outside of the range of training data. Newer architectures such as attention mechanisms, transformers, and graph neural networks are capable of learning complicated spatiotemporal interactions. Recent hybrid schemes between mechanistic multiphase flow models and data-driven corrections are promising directions to the trade-off between accuracy and physical consistency.

5.6.2 Dataset Expansion

The robustness of the model could be improved by the inclusion of a larger dataset that would include a wider geographical area, new formations, a wide range of fluid systems, various well setups, both horizontal and deviated wells, and a wider range of operational parameters. The integration of time-series data would facilitate analysis and prediction of transient effects and more than instantaneous prediction of pressure.

5.6.3 Field Testing and Validation

The generalizability would be determined by the use of extended blind testing programs on various operators, basins, and types of wells. Operational readiness would be validated by real-time implementation pilots that will have live SCADA integration and real-time comparison with concurrent downhole gauge measurements. The accuracy of long-term performance monitoring tracking models would highlight performance decay and retraining needs that will guide maintenance practices.

5.6.4 Software Development

The adoption by personnel in the field would be supported by user-friendly interfaces that would abstract technical complexity. Implementation based on cloud deployment with automated data pipelines would allow the implementation to scale with large populations of wells. Standardized SCADA integration APIs would facilitate the implementation of SCADA in a wide range of operating environments.

5.6.5 Extension to Other Applications

The observed effectiveness encourages extending to horizontal and deviated wells and adding another artificial lift system (electric submersible pumps, plunger lift, or beam pumps) and other stimulation treatments (acid fracturing, matrix acidizing, or foam fracturing). Transfer learning strategies can lead to fast model development of similar applications based on the existing models.

The developed models should not be simply deployed outside the Western Desert of Egypt without proper validation and adaptation. Prior to deployment in a different geographical basin, the operators should first compare the feature distributions between the operational parameters in the new basin and the ranges of the training data to detect any potential risks of extrapolating the model. In the case of the new basin's parameters being within the bounds of the training data, we recommend direct deployment accompanied by simultaneous monitoring and periodic downhole gauge measurements for a period of validation to assure the prediction accuracy before full-scale gauge monitoring is switched off. In the case of significant distributional changes, transfer learning is the most computationally economical approach for adaptation in which the architecture of the existing model is preserved and only the last few layers are retrained with a small representative sample (only 10-20 wells) from the new basin rather than training a new model from scratch. A fixed time-based retraining cycle (9-12 months) is proposed as a nominal maintenance policy, while ongoing statistical monitoring of the model inputs based on techniques such as the Kolmogorov-Smirnov test is used to detect operational drift, triggering retraining as soon as the model begins to drift and prediction quality starts to suffer.

Developing the proposed framework for deviated and horizontal wells will be an important and technically challenging area of research that requires significant rethinking of the data structure rather than just including inclination angle as an additional scalar input. In vertical wells, depth is a complete description of wellbore geometry. In deviated and horizontal wells, geometry is described by a directional survey, which includes a sequence of multiple rows of measured depth, true vertical depth, and inclination angle that together represent the entire three-dimensional wellbore trajectory. To tackle this structural complexity, a neural network with two branches is proposed. The first branch of the network transforms the sequence of the directional survey through a recurrent neural network with a long short-term memory unit that processes the survey row by row and outputs a wellbore geometry embedding vector of a fixed size that contains information about the entire spatial wellbore trajectory regardless of the number of stations in the survey. The second branch processes the scalar operational parameters through dense layers, which are the same as the neural networks presented in this study. The outputs of both branches are then combined at a concatenation layer whose output is fed into final prediction layers that output bottomhole pressure. This design explicitly disentangles the well geometry from the operational parameters and enables the model to learn the physics of how they interact to determine bottomhole pressure. The LSTM branch will need a dataset of at least 200 to 300 wells with varied wellbore geometry profiles to learn embedding vectors that are representative of the wellbore geometry. A future research priority should be the collection of large, diverse horizontal well data sets from various basins to enable the full dual-branch implementation and benchmark against traditional multiphase flow correlations for deviated wellbore geometry.

NOMENCLATURE

Abbreviations and Acronyms

ADAB – Adaptive Boosting
AI – Artificial Intelligence
API – American Petroleum Institute
BHP – Bottomhole Pressure
BHP-CTD – Bottomhole Pressure at Coiled Tubing Depth
CB – CatBoost
CTD – Coiled Tubing Depth
DT – Decision Tree
EUE – External Upset End
FFR – Fluid Flow Rate
FFR-S – Surface Fluid Flow Rate
GB – Gradient Boosting
GB-CB – Gradient Boosting (CatBoost)
GB-SKL – Gradient Boosting (scikit-learn)
GL – Gel Load
GLIR – Gas Lift Injection Rate
GOR – Gas–Oil Ratio
GP-SR – Genetic Programming Symbolic Regression
HPG – Hydroxypropyl Guar
HSP – High-Strength Proppant
ID – Inside Diameter
IPD – Injection Point Depth
IPR – Inflow Performance Relationship
IQR – Interquartile Range
ISP – Intermediate-Strength Proppant
kNN – k-Nearest Neighbors
L-BFGS – Limited-memory Broyden–Fletcher–Goldfarb–Shanno
LR – Linear Regression
MAE – Mean Absolute Error
MAPE – Mean Absolute Percentage Error
ML – Machine Learning
MLP – Multilayer Perceptron
MSE – Mean Squared Error
NN – Neural Network
NR – Nitrogen Rate
OD – Outside Diameter
OG – Oil Gravity
OGIP – Operating Gas Injection Pressure
PD – Perforation Depth
PGD – Pressure Gauge Depth
PSG – Proppant Specific Gravity
PZ – Proppant Size

RF – Random Forest
RMSE – Root Mean Squared Error
RRSCV – Repeated Random Sampling Cross-Validation
RT – Reservoir Temperature
SCADA – Supervisory Control and Data Acquisition
SGD – Stochastic Gradient Descent
SHAP – SHapley Additive exPlanations
SP – Surface Pressure
SPC – Surface Proppant Concentration
SPE – Society of Petroleum Engineers
SVM – Support Vector Machine
SVR – Support Vector Regression
TD – Tubing Depth
TID – Tubing Inside Diameter
TUFFP – Tulsa University Fluid Flow Projects
WC – Water Cut
WHP – Wellhead Pressure
WHT – Wellhead Temperature
XGB – Extreme Gradient Boosting

Greek Symbols

α – Regularization parameter
ε – Regression loss tolerance
η – Learning rate
λ – L2 regularization parameter
ρ – Spearman's rank correlation coefficient (–)
ρ_f – Fluid density (g/cm^3)
σ – Standard deviation
Δp – Pressure drop (psi)

Latin Symbols

A – Minimum dataset value (normalization)
B – Maximum dataset value (normalization)
C – SVM penalty parameter (–)
C_v – Volume fraction of solids (–)
d – Pipe inside diameter (in.)
d_i – Rank difference
k – Number of nearest neighbors
k′ – Fluid consistency index (lb·s$^{n'}$/ft^2)
L – Pipe length (ft)
n – Number of data points
n′ – Flow behavior index
P – Pressure (psia)
Q – Pump rate (bbl/min)

r – Pearson correlation coefficient (–)
Re – Reynolds number (–)
t – Time (min)
T – Temperature (°F)
X – Variable
$\bar{X}$ – Mean of X
X_{max} – Maximum value of X
X_{min} – Minimum value of X
X_{norm} – Normalized value
Y – Target variable
$\bar{Y}$ – Mean of Y

Units of Measurement

°API – Degrees API Gravity
°F – Degrees Fahrenheit
bbl/min – Barrels per minute
ft – Feet
g/cm^3 – Grams per cubic centimeter
in. – Inches
lb/gal – Pounds per gallon
lb/mgal – Pounds per thousand gallons
mmscf/d – Million standard cubic feet per day
ppm – Parts per million
psi – Pounds per square inch
psia – Pounds per square inch absolute
scf/m – Standard cubic feet per minute
scf/stb – Standard cubic feet per stock tank barrel
stb/d – Stock tank barrels per day

Subscripts and Superscripts

i – Data index
j – Feature index
k – Iteration index
max – Maximum
min – Minimum
norm – Normalized value
n′ – Power-law exponent

References

(1) Yue, P.; Yang, H.; He, C.; Yu, G. M.; Sheng, J. J.; Guo, Z. L.; Guo, C. Q.; Chen, X. F. Theoretical Approach for the Calculation of the Pressure Drop in a Multibranch Horizontal Well with Variable Mass Transfer. *ACS Omega* **2020**, *5* (45), 29209–29221. https://doi.org/10.1021/acsomega.0c03971.

(2) Akinsete, O.; Adesiji, B. Improved Mechanistic and Intelligent Models for Bottom-Hole Pressure from Vertical Oil Wellhead Data. *PSE* **2025**, *9* (2), 111–119. https://doi.org/10.11648/j.pse.20250902.16.

(3) Abd El-Moniem, M. A.; El-Banbi, A. H. Expert Solution for Effects of Input Parameters on Multiphase Flow Correlations. *Journal of Engineering Research* **2021**, *9* (2), 336–359. https://doi.org/10.36909/jer.v9i2.9067.

(4) Liang, T.; Deng, Z.; Wu, J.; Xu, F.; Zheng, L.; Yang, M.; Zhou, F. A Power-Law-Based Predictive Model for Proppant Settling Velocity in Non-Newtonian Fluid. *Processes* **2025**, *13* (8), 2631. https://doi.org/10.3390/pr13082631.

(5) Lu, Q.; Pal, R. Steady Shear Rheology and Surface Activity of Polymer-Surfactant Mixtures. *Polymers* **2025**, *17* (3), 364. https://doi.org/10.3390/polym17030364.

(6) Bai, Z.; Li, M. Prediction of Proppant Settling Velocity in Fiber-Containing Fracturing Fluids. *ACS Omega* **2023**, *8* (35), 31857–31869. https://doi.org/10.1021/acsomega.3c03369.

(7) Rodriguez, O. M. H.; Oliemans, R. V. A. Experimental Study on Oil–Water Flow in Horizontal and Slightly Inclined Pipes. *International Journal of Multiphase Flow* **2006**, *32* (3), 323–343. https://doi.org/10.1016/j.ijmultiphaseflow.2005.11.001.

(8) Moore, T. A.; Friederich, M. C. Defining Uncertainty: Comparing Resource/Reserve Classification Systems for Coal and Coal Seam Gas. *Energies* **2021**, *14* (19), 6245. https://doi.org/10.3390/en14196245.

(9) Raab, T. .; Reinsch, T. .; Aldaz Cifuentes, S. R.; Henninges, J. . Real-Time Well-Integrity Monitoring Using Fiber-Optic Distributed Acoustic Sensing. *SPE Journal* **2019**, *24* (05), 1997–2009. https://doi.org/10.2118/195678-PA.

(10) Liu, Y.; Wu, N.; Luo, C.; Tian, W.; Li, N.; Cao, G.; Wang, Q.; Gao, C.; Deng, Q. A New Criterion for Predicting Liquid Loading in Vertical Gas Wells. *Results in Engineering* **2026**, *29*, 108909. https://doi.org/10.1016/j.rineng.2025.108909.

(11) Enyekwe, A. E.; Ajienka, J. A. Comparative Analysis of Permanent Downhole Gauges and Their Applications. In *SPE Nigeria Annual International Conference and Exhibition*; SPE: Lagos, Nigeria, 2014; p SPE-172435-MS. https://doi.org/10.2118/172435-MS.

(12) Carpenter, C. Study Reviews Downhole Wireless Technologies and Improvements. *Journal of Petroleum Technology* **2022**, *74* (03), 70–72. https://doi.org/10.2118/0322-0070-JPT.

(13) Pourafshary, P.; Podio, A. L.; Sepehrnoori, K. The Effect of Pressure Gauge Placement on Well Testing Analysis Accuracy. *Petroleum Science and Technology* **2009**, *27* (6), 625–642. https://doi.org/10.1080/10916460701857748.

(14) Matsumoto, N.; Sudo, Y.; Sinha, B. K.; Niwa, M. Long-Term Stability and Performance Characteristics of Crystal Quartz Gauge at High Pressures and Temperatures. *IEEE Trans. Ultrason., Ferroelect., Freq. Contr.* **2000**, *47* (2), 346–354. https://doi.org/10.1109/58.827419.

(15) Clowes, J.; Edwards, J.; Grudinin, I.; Kluth, E. L. E.; Varnham, M. P.; Zervas, M. N.; Crawley, C. M.; Kutlik, R. L. Low Drift Fibre Optic Pressure Sensor for Oil Fielddownhole Monitoring. *Electron. Lett.* **1999**, *35* (11), 926–927. https://doi.org/10.1049/el:19990646.

(16) Aminzadeh, F. Hydraulic Fracturing, An Overview. *j sustain energy engng* **2018**, *6* (3), 204–228. https://doi.org/10.7569/JSEE.2018.629512.

(17) Zhao, M. Field Experiments and Main Understanding of Shale Oil Hydraulic Fracturing. *Front. Earth Sci.* **2024**, *12*, 1410524. https://doi.org/10.3389/feart.2024.1410524.
(18) Hyman, J. D.; Jiménez-Martínez, J.; Viswanathan, H. S.; Carey, J. W.; Porter, M. L.; Rougier, E.; Karra, S.; Kang, Q.; Frash, L.; Chen, L.; Lei, Z.; O'Malley, D.; Makedonska, N. Understanding Hydraulic Fracturing: A Multi-Scale Problem. *Phil. Trans. R. Soc. A.* **2016**, *374* (2078), 20150426. https://doi.org/10.1098/rsta.2015.0426.
(19) Department of Energy, University of Sao Paulo, Brazil; Rivera, Dr. J. C. HYDRAULIC FRACTURING IN OIL AND GAS WELLS: TECHNIQUES, INNOVATION, AND ENVIRONMENTAL IMPACTS. *ijnget* **2025**, *2* (1), 8–14. https://doi.org/10.55640/ijnget-v02i01-02.
(20) Zhai, L.; Xun, Y.; Liu, H.; Qi, B.; Wu, J.; Wang, Y.; Chen, C. An Investigation of Hydraulic Fracturing Initiation and Location of Hydraulic Fracture in Preforated Oil Shale Formations. *Sci Rep* **2025**, *15* (1), 4196. https://doi.org/10.1038/s41598-025-88774-y.
(21) Dong, J.; Qu, H.; Zhang, J.; Han, F.; Zhou, F.; Shi, P.; Shi, J.; Yu, T. Numerical Simulation of Hydraulic Fracture Propagation under Energy Supplement Conditions. *Front. Earth Sci.* **2023**, *11*, 1269159. https://doi.org/10.3389/feart.2023.1269159.
(22) Yang, P.; Zhang, S.; Zou, Y.; Li, J.; Ma, X.; Tian, G.; Wang, J. Fracture Propagation, Proppant Transport and Parameter Optimization of Multi-Well Pad Fracturing Treatment. *Petroleum Exploration and Development* **2023**, *50* (5), 1225–1235. https://doi.org/10.1016/S1876-3804(23)60461-6.
(23) Ai, K.; Duan, L.; Gao, H.; Jia, G. Hydraulic Fracturing Treatment Optimization for Low Permeability Reservoirs Based on Unified Fracture Design. *Energies* **2018**, *11* (7), 1720. https://doi.org/10.3390/en11071720.
(24) Liu, G.; Han, C.; Yang, H.; Xiu, J.; Li, X.; Hao, Z.; Wei, B.; Lv, N. Main Control Factors of Fracture Propagation in Reservoir: A Review. *ACS Omega* **2024**, *9* (1), 117–136. https://doi.org/10.1021/acsomega.3c08547.
(25) Shi, W.; Guo, M.; Huang, Z.; Zhang, Z.; Zhang, C.; Shi, Y. Study on Development of Shale Gas Horizontal Well With Time-Phased Staged Fracturing and Refracturing: Follow-Up and Evaluation of Well R9-2, A Pilot Well in Fuling Shale Gas Field. *IEEE Access* **2021**, *9*, 117027–117039. https://doi.org/10.1109/ACCESS.2021.3105186.
(26) Prabhakaran, R.; De Pater, H.; Shaoul, J. Pore Pressure Effects on Fracture Net Pressure and Hydraulic Fracture Containment: Insights from an Empirical and Simulation Approach. *Journal of Petroleum Science and Engineering* **2017**, *157*, 724–736. https://doi.org/10.1016/j.petrol.2017.07.009.
(27) Guo, Y.; Zhang, M.; Yang, H.; Wang, D.; Ramos, M. A.; Hu, T. S.; Xu, Q. Friction Challenge in Hydraulic Fracturing. *Lubricants* **2022**, *10* (2), 14. https://doi.org/10.3390/lubricants10020014.
(28) Shah, S. N.; Lee, Y. N. Friction Pressures of Proppant-Laden Hydraulic Fracturing Fluids. *SPE Production Engineering* **1986**, *1* (06), 437–445. https://doi.org/10.2118/13836-PA.
(29) Wang, Q.; Yang, Z.; Zeng, L.; Du, A.; Chen, Y.; Luo, W. Research on the Optimization of Continuous Gas Lift Production from Multiple Wells on the Platform. *Processes* **2025**, *13* (2), 478. https://doi.org/10.3390/pr13020478.
(30) Khamehchi, E.; Rashidi, F.; Omranpour, H.; Ghidary, S. S.; Ebrahimian, A.; Rasouli, H. Intelligent System for Continuous Gas Lift Operation and Design with Unlimited Gas Supply. *J. of Applied Sciences* **2009**, *9* (10), 1889–1897. https://doi.org/10.3923/jas.2009.1889.1897.
(31) Abu-Bakri, J.; Jafari, A.; Namdar, H.; Ahmadi, G. Increasing Productivity by Using Smart Gas for Optimal Management of the Gas Lift Process in a Cluster of Wells. *Sci Rep* **2024**, *14* (1), 15489. https://doi.org/10.1038/s41598-024-63506-w.
(32) Moffett, R. E.; Seale, S. R. Real Gas Lift Optimization: An Alternative to Timer Based Intermittent Gas Lift. In *Abu Dhabi International Petroleum Exhibition & Conference*; SPE: Abu Dhabi, UAE, 2017; p D021S048R004. https://doi.org/10.2118/188480-MS.

(33) Zhong, H.; Zhao, C.; Xu, Z.; Zheng, C. Economical Optimum Gas Allocation Model Considering Different Types of Gas-Lift Performance Curves. *Energies* **2022**, *15* (19), 6950. https://doi.org/10.3390/en15196950.
(34) Khoshkbarchi, M.; Rahmanian, M.; Cordazzo, J.; Nghiem, L. Application of Mesh Adaptive Derivative-Free Optimization Technique for Gas-Lift Optimization in an Integrated Reservoirs, Wells, and Facilities Modeling Environment. In *SPE Canada Heavy Oil Conference*; SPE: Virtual, 2020; p D041S008R003. https://doi.org/10.2118/199923-MS.
(35) Suharto, M. A.; Risal, A. R.; Subandi, A. N.; Trjangganung, K.; Zain, A. M.; Ahnap, M. S. Maximizing Oil Production by Leveraging Python for Gas Lift Optimization Through Well Modelling. In *International Petroleum Technology Conference*; IPTC: Kuala Lumpur, Malaysia, 2025; p D032S010R016. https://doi.org/10.2523/IPTC-25032-EA.
(36) Mammadov, A. V.; Sultanova, A. V.; Mammadov, R. M. Grouping of Gas Lift Wells Based on Their Interaction. In *SPE Caspian Technical Conference and Exhibition*; SPE: Baku, Azerbaijan, 2023; p D021S006R008. https://doi.org/10.2118/217515-MS.
(37) Elwan, M.; Davani, E.; Ghedan, S. G.; Mousa, H.; Kansao, R.; Surendra, M.; Deng, L.; Korish, M.; Shahin, E.; Ibrahim, M.; Eid, T.; Rouis, L. Automated Well Production and Gas Lift System Diagnostics and Optimization Using Augmented AI Approach in a Complex Offshore Field. In *SPE Annual Technical Conference and Exhibition*; SPE: Dubai, UAE, 2021; p D011S013R001. https://doi.org/10.2118/206213-MS.
(38) Jin, M.; Emami-Meybodi, H.; Ahmadi, M. Flowing Bottomhole Pressure during Gas Lift in Unconventional Oil Wells. *SPE Journal* **2024**, 1–13. https://doi.org/10.2118/214832-PA.
(39) Baryshnikov, E. S.; Kanin, E. A.; Osiptsov, A. A.; Vainshtein, A. L.; Burnaev, E. V.; Paderin, G. V.; Prutsakov, A. S.; Ternovenko, S. O. Adaptation of Steady-State Well Flow Model on Field Data for Calculating the Flowing Bottomhole Pressure. In *SPE Russian Petroleum Technology Conference*; SPE: Virtual, 2020; p D023S007R008. https://doi.org/10.2118/201942-MS.
(40) Cox, S. A.; Sutton, R. P.; Blasingame, T. A. Errors Introduced by Multiphase Flow Correlations on Production Analysis. In *SPE Annual Technical Conference and Exhibition*; SPE: San Antonio, Texas, USA, 2006; p SPE-102488-MS. https://doi.org/10.2118/102488-MS.
(41) Obeida, T. A.; Mosallam, Y. H.; Al Mehari, Y. S. Calculation of Flowing Bottomhole Pressure Constraint Based on Bubblepoint-Pressure-vs.-Depth Relationship. In *SPE Middle East Oil and Gas Show and Conference*; SPE: Manama, Bahrain, 2007; p SPE-104985-MS. https://doi.org/10.2118/104985-MS.
(42) Usov, E. V.; Ulyanov, V. N.; Butov, A. A.; Chuhno, V. I.; Lyhin, P. A. Modelling Multiphase Flows of Hydrocarbons in Gas-Condensate and Oil Wells. *Math Models Comput Simul* **2020**, *12* (6), 1005–1013. https://doi.org/10.1134/S2070048220060162.
(43) Chanchlani, K.; Sukanandan, J.; Devshali, S.; Kumar, V.; Kumar, A. Delineating and Decoding the Qualitative Interpretation of Gas Lift Instabilities in the Continuous Gas Lift Wells of Mehsana. In *International Petroleum Technology Conference*; IPTC: Kuala Lumpur, Malaysia, 2025; p D022S006R008. https://doi.org/10.2523/IPTC-24756-MS.
(44) Hashmi, U. I.; Osisanya, S.; Rahman, M.; Ghosh, B.; Subaihi, M. H.; Kuwair, H. F. B.; Fathy, A.; Afagwu, C.; Mahmoud, M. Optimization of Nitrogen Lifting Operation in Horizontal Wells of Carbonate Formation. In *International Petroleum Technology Conference*; IPTC: Riyadh, Saudi Arabia, 2022; p D031S096R001. https://doi.org/10.2523/IPTC-22397-MS.
(45) Aitken, B.; Livescu, S.; Craig, S. Coiled Tubing Software Models and Field Applications – A Review. *Journal of Petroleum Science and Engineering* **2019**, *182*, 106308. https://doi.org/10.1016/j.petrol.2019.106308.

(46) Han, G.; Ma, G.; Gao, Y.; Zhang, H.; Ling, K. A New Transient Model to Simulate and Optimize Liquid Unloading with Coiled Tubing Conveyed Gas Lift. *Journal of Petroleum Science and Engineering* **2021**, *200*, 108394. https://doi.org/10.1016/j.petrol.2021.108394.
(47) Michel, G.; Civan, F. Modeling Rapid Multiphase Flow in Wells and Pipelines Under Nonequilibrium and Nonisothermal Conditions. In *Rocky Mountain Oil & Gas Technology Symposium*; SPE: Denver, Colorado, U.S.A., 2007; p SPE-107958-MS. https://doi.org/10.2118/107958-MS.
(48) Yudin, E.; Khabibullin, R.; Smirnov, N.; Vodopyan, A.; Goridko, K.; Chigarev, G.; Zamakhov, S. New Applications of Transient Multiphase Flow Models in Wells and Pipelines for Production Management. In *SPE Russian Petroleum Technology Conference*; SPE: Virtual, 2020; p D013S027R004. https://doi.org/10.2118/201884-MS.
(49) Tyagi, A.; Gupta, A. Machine Learning Based Prediction of Pressure Drop, Liquid-Holdup and Flow Pattern in Multiphase Flows. In *SPE EuropEC - Europe Energy Conference featured at the 84th EAGE Annual Conference & Exhibition*; SPE: Vienna, Austria, 2023; p D021S002R002. https://doi.org/10.2118/214348-MS.
(50) Cremaschi, S.; Kouba, G. E.; Subramani, H. J. Characterization of Confidence in Multiphase Flow Predictions. *Energy Fuels* **2012**, *26* (7), 4034–4045. https://doi.org/10.1021/ef300190p.
(51) Szalinski, L.; Abdulkareem, L. A.; Da Silva, M. J.; Thiele, S.; Beyer, M.; Lucas, D.; Hernandez Perez, V.; Hampel, U.; Azzopardi, B. J. Comparative Study of Gas–Oil and Gas–Water Two-Phase Flow in a Vertical Pipe. *Chemical Engineering Science* **2010**, *65* (12), 3836–3848. https://doi.org/10.1016/j.ces.2010.03.024.
(52) Al-Qasim, A. S.; Almudairis, F.; Bin Omar, A.; Omair, A. Optimizing Production Facilities Using a Transient Multiphase Flow Simulator. In *Volume 8: Polar and Arctic Sciences and Technology; Petroleum Technology*; American Society of Mechanical Engineers: Glasgow, Scotland, UK, 2019; p V008T11A048. https://doi.org/10.1115/OMAE2019-95002.
(53) Jacobs, T. The New Pathways of Multiphase Flow Modeling. *Journal of Petroleum Technology* **2015**, *67* (02), 62–68. https://doi.org/10.2118/0215-0062-JPT.
(54) Danielson, T. J. Transient Multiphase Flow: Past, Present, and Future with Flow Assurance Perspective. *Energy Fuels* **2012**, *26* (7), 4137–4144. https://doi.org/10.1021/ef300300u.
(55) Ye, J.; Guo, L. Multiphase Flow Pattern Recognition in Pipeline–Riser System by Statistical Feature Clustering of Pressure Fluctuations. *Chemical Engineering Science* **2013**, *102*, 486–501. https://doi.org/10.1016/j.ces.2013.08.048.
(56) Abubakar, A.; Al-Wahaibi, Y.; Al-Wahaibi, T.; Al-Hashmi, A.; Al-Ajmi, A.; Eshrati, M. Effect of Low Interfacial Tension on Flow Patterns, Pressure Gradients and Holdups of Medium-Viscosity Oil/Water Flow in Horizontal Pipe. *Experimental Thermal and Fluid Science* **2015**, *68*, 58–67. https://doi.org/10.1016/j.expthermflusci.2015.02.017.
(57) Giraudeau, M.; Mureithi, N. W.; Pettigrew, M. J. Two-Phase Flow-Induced Forces on Piping in Vertical Upward Flow: Excitation Mechanisms and Correlation Models. *Journal of Pressure Vessel Technology* **2013**, *135* (3), 030907. https://doi.org/10.1115/1.4024210.
(58) Zou, S.; Guo, L.; Xie, C. Fast Recognition of Global Flow Regime in Pipeline-Riser System by Spatial Correlation of Differential Pressures. *International Journal of Multiphase Flow* **2017**, *88*, 222–237. https://doi.org/10.1016/j.ijmultiphaseflow.2016.08.007.
(59) Ahmadi, M. A.; Galedarzadeh, M.; Shadizadeh, S. R. Low Parameter Model to Monitor Bottom Hole Pressure in Vertical Multiphase Flow in Oil Production Wells. *Petroleum* **2016**, *2* (3), 258–266. https://doi.org/10.1016/j.petlm.2015.08.001.
(60) Kaji, R.; Azzopardi, B. J. The Effect of Pipe Diameter on the Structure of Gas/Liquid Flow in Vertical Pipes. *International Journal of Multiphase Flow* **2010**, *36* (4), 303–313. https://doi.org/10.1016/j.ijmultiphaseflow.2009.11.010.

(61) Zhang, H.-Q.; Sarica, C.; Pereyra, E. Review of High-Viscosity Oil Multiphase Pipe Flow. *Energy Fuels* **2012**, *26* (7), 3979–3985. https://doi.org/10.1021/ef300179s.
(62) Firouzi, M.; Towler, B.; Rufford, T. E. Developing New Mechanistic Models for Predicting Pressure Gradient in Coal Bed Methane Wells. *Journal of Natural Gas Science and Engineering* **2016**, *33*, 961–972. https://doi.org/10.1016/j.jngse.2016.04.035.
(63) Hasan, A. R.; Kabir, C. S.; Sayarpour, M. Simplified Two-Phase Flow Modeling in Wellbores. *Journal of Petroleum Science and Engineering* **2010**, *72* (1–2), 42–49. https://doi.org/10.1016/j.petrol.2010.02.007.
(64) Thome, J. R.; Bar-Cohen, A.; Revellin, R.; Zun, I. Unified Mechanistic Multiscale Mapping of Two-Phase Flow Patterns in Microchannels. *Experimental Thermal and Fluid Science* **2013**, *44*, 1–22. https://doi.org/10.1016/j.expthermflusci.2012.09.012.
(65) Sarica, C.; Zhang, H.-Q.; Wilkens, R. J. Sensitivity of Slug Flow Mechanistic Models on Slug Length. *Journal of Energy Resources Technology* **2011**, *133* (4), 043001. https://doi.org/10.1115/1.4005242.
(66) Abdulwarith, A.; Ammar, M.; Kakadjian, S.; McLaughlin, N.; Dindoruk, B. A Hybrid Physics Augmented Predictive Model for Friction Pressure Loss in Hydraulic Fracturing Process Based on Experimental and Field Data. In *Offshore Technology Conference*; OTC: Houston, Texas, USA, 2024; p D031S030R006. https://doi.org/10.4043/35310-MS.
(67) Zendehboudi, S.; Rezaei, N.; Lohi, A. Applications of Hybrid Models in Chemical, Petroleum, and Energy Systems: A Systematic Review. *Applied Energy* **2018**, *228*, 2539–2566. https://doi.org/10.1016/j.apenergy.2018.06.051.
(68) Woldesemayat, M. A.; Ghajar, A. J. Comparison of Void Fraction Correlations for Different Flow Patterns in Horizontal and Upward Inclined Pipes. *International Journal of Multiphase Flow* **2007**, *33* (4), 347–370. https://doi.org/10.1016/j.ijmultiphaseflow.2006.09.004.
(69) Abdul-Majeed, G. H.; Al-Mashat, A. M. A Unified Correlation for Predicting Slug Liquid Holdup in Viscous Two-Phase Flow for Pipe Inclination from Horizontal to Vertical. *SN Appl. Sci.* **2019**, *1* (1), 71. https://doi.org/10.1007/s42452-018-0081-0.
(70) Barnea, D.; Roitberg, E.; Shemer, L. Spatial Distribution of Void Fraction in the Liquid Slug in the Whole Range of Pipe Inclinations. *International Journal of Multiphase Flow* **2013**, *52*, 92–101. https://doi.org/10.1016/j.ijmultiphaseflow.2012.12.005.
(71) Abdul-Majeed, G.; Al-Sarkhi, A.; Al-Fatlawi, O. F.; Mohmmed, A. O. Empirical Model for Predicting Slug-Pseudo Slug and Slug-Churn Transitions of Upward Air/Water Flow. *Geoenergy Science and Engineering* **2025**, *246*, 213613. https://doi.org/10.1016/j.geoen.2024.213613.
(72) Pedersen, S.; Durdevic, P.; Yang, Z. Challenges in Slug Modeling and Control for Offshore Oil and Gas Productions: A Review Study. *International Journal of Multiphase Flow* **2017**, *88*, 270–284. https://doi.org/10.1016/j.ijmultiphaseflow.2016.07.018.
(73) Liang, C.; Xiong, W.; Wang, H.; Wang, Z. Experimental and OLGA Modeling Investigation for Slugging in Underwater Compressed Gas Energy Storage Systems. *Applied Sciences* **2023**, *13* (17), 9575. https://doi.org/10.3390/app13179575.
(74) Berna, C.; Escrivá, A.; Muñoz-Cobo, J. L.; Herranz, L. E. Development of New Correlations for Annular Flow; València, Spain, 2015; pp 451–462. https://doi.org/10.2495/MPF150381.
(75) Ye, J.; Do, N. C.; Zeng, W.; Lambert, M. Physics-Informed Neural Networks for Hydraulic Transient Analysis in Pipeline Systems. *Water Research* **2022**, *221*, 118828. https://doi.org/10.1016/j.watres.2022.118828.
(76) Asimiea, N. W.; Ebere, E. N. Using Machine Learning to Predict Permeability from Well Logs: A Comparative Study of Different Models. In *SPE Nigeria Annual International Conference and Exhibition*; SPE: Lagos, Nigeria, 2023; p D021S004R005. https://doi.org/10.2118/217158-MS.

(77) Safarov, A.; Iskandarov, V.; Solomonov, D. Application of Machine Learning Techniques for Rate of Penetration Prediction. In *SPE Annual Caspian Technical Conference*; SPE: Nur-Sultan, Kazakhstan, 2022; p D021S013R002. https://doi.org/10.2118/212088-MS.
(78) Fan, Z.; Chen, J.; Zhang, T.; Shi, N.; Zhang, W. Machine Learning for Formation Tightness Prediction and Mobility Prediction. In *SPE Annual Technical Conference and Exhibition*; SPE: Dubai, UAE, 2021; p D011S018R005. https://doi.org/10.2118/206208-MS.
(79) Mukherjee, T.; Burgett, T.; Ghanchi, T.; Donegan, C.; Ward, T. Predicting Gas Production Using Machine Learning Methods: A Case Study. In *SEG Technical Program Expanded Abstracts 2019*; Society of Exploration Geophysicists: San Antonio, Texas, 2019; pp 2248–2252. https://doi.org/10.1190/segam2019-3215692.1.
(80) Al Shehri, F. H.; Gryzlov, A.; Al Tayyar, T.; Arsalan, M. Utilizing Machine Learning Methods to Estimate Flowing Bottom-Hole Pressure in Unconventional Gas Condensate Tight Sand Fractured Wells in Saudi Arabia. In *SPE Russian Petroleum Technology Conference*; SPE: Virtual, 2020; p D043S032R002. https://doi.org/10.2118/201939-MS.
(81) Gabry, M. A.; Ali, A. G.; Elsawy, M. S. Application of Machine Learning Model for Estimating the Geomechanical Rock Properties Using Conventional Well Logging Data. In *Offshore Technology Conference*; OTC: Houston, Texas, USA, 2023; p D021S028R004. https://doi.org/10.4043/32328-MS.
(82) El Khouly, I.; Sabet, A.; El-Fattah, M. A. A.; Bulatnikov, M. Integration Between Different Hydraulic Fracturing Techniques and Machine Learning in Optimizing and Evaluating Hydraulic Fracturing Treatment. In *International Petroleum Technology Conference*; IPTC: Dhahran, Saudi Arabia, 2024; p D021S084R003. https://doi.org/10.2523/IPTC-24296-MS.
(83) Zhuang, X.; Liu, Y.; Hu, Y.; Guo, H.; Nguyen, B. H. Prediction of Rock Fracture Pressure in Hydraulic Fracturing with Interpretable Machine Learning and Mechanical Specific Energy Theory. *Rock Mechanics Bulletin* **2025**, *4* (2), 100173. https://doi.org/10.1016/j.rockmb.2024.100173.
(84) Yang, C.; Xu, C.; Ma, Y.; Qu, B.; Liang, Y.; Xu, Y.; Xiao, L.; Sheng, Z.; Fan, Z.; Zhang, X. A Novel Paradigm for Parameter Optimization of Hydraulic Fracturing Using Machine Learning and Large Language Model. *ijacsa* **2025**, *16* (3). https://doi.org/10.14569/IJACSA.2025.01603108.
(85) Gharieb, A.; Gabry, M. A.; Soliman, M. Y. The Role of Personalized Generative AI in Advancing Petroleum Engineering and Energy Industry: A Roadmap to Secure and Cost-Efficient Knowledge Integration: A Case Study. In *SPE Annual Technical Conference and Exhibition*; SPE: New Orleans, Louisiana, USA, 2024; p D011S007R002. https://doi.org/10.2118/220716-MS.
(86) Hegde, C.; Daigle, H.; Millwater, H.; Gray, K. Analysis of Rate of Penetration (ROP) Prediction in Drilling Using Physics-Based and Data-Driven Models. *Journal of Petroleum Science and Engineering* **2017**, *159*, 295–306. https://doi.org/10.1016/j.petrol.2017.09.020.
(87) Brantson, E. T.; Adu-Awuku, J.; Asase, W.; Obeng, S. D. A.; Kwakye-Tannor, M. B.; Nuamah, J. A.; Aboagye, E.; Appiah-Adjei, F.; Nuetor, F. D.; Opoku, P. Optimization of Well Trajectory with Machine Learning Algorithms for Geosteering Directional Drilling. *Journal of Geophysics and Engineering* **2025**, *22* (5), 1245–1265. https://doi.org/10.1093/jge/gxaf055.
(88) Zuo, C.; Li, Z.; Dai, Z.; Wang, X.; Wang, Y. A Pattern Classification Distribution Method for Geostatistical Modeling Evaluation and Uncertainty Quantification. *Remote Sensing* **2023**, *15* (11), 2708. https://doi.org/10.3390/rs15112708.
(89) Halotel, J.; Demyanov, V.; Gardiner, A. Value of Geologically Derived Features in Machine Learning Facies Classification. *Math Geosci* **2020**, *52* (1), 5–29. https://doi.org/10.1007/s11004-019-09838-0.
(90) Yehia, T. A.; Freestone, B.; Nair, G. G.; Falola, Y.; Ailneni, R. R.; Rathore, P. S.; Toms, J.; Meehan, D. N. An Ensemble-Based Automated Workflow for Decline Curve Analysis and Probabilistic Estimated Ultimate Recovery Estimation in Shale Gas Wells. *SPE Journal* **2025**, *30* (11), 6638–6655. https://doi.org/10.2118/230302-PA.

(91) Wang, X.; Zhuang, Y.; Xie, Y.; Chen, L.; Yu, W.; Li, M.; Wu, Y. Multi-Objective Optimization of Sucker Rod Pump Operating Parameters for Efficiency and Pump Life Improvement Based on Random Forest and CMA-ES. *Processes* **2025**, *13* (12), 3871. https://doi.org/10.3390/pr13123871.
(92) Zhu, H.; Yu, H.; Sun, Q.; Wang, Q.; Jing, H.; Abdikadyrov, R. Deep Learning-Based Performance Prediction of Electric Submersible Pumps Under Viscous and Gas–Liquid Flow Conditions. *Machines* **2025**, *13* (2), 135. https://doi.org/10.3390/machines13020135.
(93) Gou, L.; Yang, Z.; Min, C.; Yi, D.; Yi, L.; Li, X. A Novel Hybrid Deep Learning Methodology for Real-Time Wellhead Pressure Forecasting and Risk Warning during Shale Gas Hydraulic Fracturing. *Reliability Engineering & System Safety* **2025**, *264*, 111309. https://doi.org/10.1016/j.ress.2025.111309.
(94) Li, W.; Zhang, T.; Liu, X.; Dong, Z.; Dong, G.; Qian, S.; Yang, Z.; Zou, L.; Lin, K.; Zhang, T. Machine Learning-Based Fracturing Parameter Optimization for Horizontal Wells in Panke Field Shale Oil. *Sci Rep* **2024**, *14* (1), 6046. https://doi.org/10.1038/s41598-024-56660-8.
(95) De Rezende Faria, R.; Capron, B. D. O.; Secchi, A. R.; De Souza, M. B. Gas-Lift Optimization Using Physics-Informed Deep Reinforcement Learning. *Ind. Eng. Chem. Res.* **2024**, *63* (32), 14199–14210. https://doi.org/10.1021/acs.iecr.3c04615.
(96) Fei, X.; Ye, M.; Du, Z.; Miao, H. A Comparative Study of MLP and LSTM Neural Networks for Shale Gas Production Prediction Based on Numerical Simulation Data. *PLoS One* **2025**, *20* (11), e0336782. https://doi.org/10.1371/journal.pone.0336782.
(97) Wu, T.; Liu, Q.; Wang, Y.; Xu, Y.; Shi, J.; Yao, Y.; Chen, Q.; Liang, J.; Tang, S. Research Progress and Technology Outlook of Deep Learning in Seepage Field Prediction During Oil and Gas Field Development. *Applied Sciences* **2025**, *15* (11), 6059. https://doi.org/10.3390/app15116059.
(98) Yalamanchi, P.; Datta Gupta, S.; Upadhyay, R. Evaluating the Effectiveness of Ensemble Machine Learning Approaches for Pore Pressure Prediction Using Petrophysical Log Data in Carbonate Reservoir. *Acta Geophys.* **2025**, *73* (3), 2591–2619. https://doi.org/10.1007/s11600-025-01530-8.
(99) Kharazi Esfahani, P.; Peiro Ahmady Langeroudy, K.; Khorsand Movaghar, M. R. Enhanced Machine Learning—Ensemble Method for Estimation of Oil Formation Volume Factor at Reservoir Conditions. *Sci Rep* **2023**, *13* (1), 15199. https://doi.org/10.1038/s41598-023-42469-4.
(100) Wang, Y.; Wagner, N.; Rondinelli, J. M. Symbolic Regression in Materials Science. *MRS Communications* **2019**, *9* (3), 793–805. https://doi.org/10.1557/mrc.2019.85.
(101) Taskin, B.; Xie, W.; Lazebnik, T. Knowledge Integration for Physics-Informed Symbolic Regression Using Pre-Trained Large Language Models. *Sci Rep* **2026**, *16* (1), 1614. https://doi.org/10.1038/s41598-026-35327-6.
(102) Makke, N.; Chawla, S. Interpretable Scientific Discovery with Symbolic Regression: A Review. *Artif Intell Rev* **2024**, *57* (1), 2. https://doi.org/10.1007/s10462-023-10622-0.
(103) Praks, P.; Rasmussen, A.; Lye, K. O.; Martinovič, J.; Praksová, R.; Watson, F.; Brkić, D. Sensitivity Analysis of Parameters for Carbon Sequestration: Symbolic Regression Models Based on Open Porous Media Reservoir Simulators Predictions. *Heliyon* **2024**, *10* (22), e40044. https://doi.org/10.1016/j.heliyon.2024.e40044.
(104) Hafsa, N.; Rushd, S.; Yousuf, H. Comparative Performance of Machine-Learning and Deep-Learning Algorithms in Predicting Gas–Liquid Flow Regimes. *Processes* **2023**, *11* (1), 177. https://doi.org/10.3390/pr11010177.
(105) Fernandes, M. D. A.; Gildin, E.; Sampaio, M. A. Machine Learning-Based Virtual Sensor for Bottom-Hole Pressure Estimation in Petroleum Wells. *Eng* **2025**, *6* (11), 318. https://doi.org/10.3390/eng6110318.
(106) Akbari, A.; Ghazi, F.; Kazemzadeh, Y. Machine Learning-Based Prediction of Well Performance Parameters for Wellhead Choke Flow Optimization. *Sci Rep* **2025**, *15* (1), 44182. https://doi.org/10.1038/s41598-025-27851-8.

(107) Yahia, H.; Romary, T.; Gerbaud, L.; Figluizzi, B.; Di Meglio, F.; Menand, S.; Mahjoub, M. Combining Machine-Learning and Physics-Based Models to Mitigate Stick-Slip in Real-Time. In *IADC/SPE International Drilling Conference and Exhibition*; SPE: Galveston, Texas, USA, 2024; p D031S022R003. https://doi.org/10.2118/217750-MS.
(108) Zhang, T.; Li, H.; Meng, Z.; Yuan, Z.; Wang, M.; Li, J. DS-DW-TimesNet-Driven Early Warning for Downhole Near-Bit Torque Vibrations. *Processes* **2025**, *13* (9), 2700. https://doi.org/10.3390/pr13092700.
(109) Liu, Q.; Li, D.; Liu, S.; Liang, H.; Zhou, Y.; Zhao, C.; Liu, K.; Hui, G.; Ni, F.; Du, P.; Wang, S. A Machine Learning-Driven Framework for Real-Time Lithology Identification and Drilling Parameter Optimization. *Processes* **2026**, *14* (1), 156. https://doi.org/10.3390/pr14010156.
(110) Zhang, K.; Zhang, X. Adaptive Soft Sensor Modeling of Chemical Processes Based on an Improved Just-in-time Learning and Random Mapping Partial Least Squares. *Journal of Chemometrics* **2024**, *38* (9), e3554. https://doi.org/10.1002/cem.3554.
(111) Dong, B.; Cui, S.; Wang, X. A Data-Driven Deep Reinforcement Learning Framework for Real-Time Economic Dispatch of Microgrids Under Renewable Uncertainty. *Energies* **2026**, *19* (6), 1481. https://doi.org/10.3390/en19061481.
(112) Choubineh, A.; Chen, J.; Wood, D. A.; Coenen, F.; Ma, F. Deep Ensemble Learning for High-Dimensional Subsurface Fluid Flow Modeling. *Engineering Applications of Artificial Intelligence* **2023**, *126*, 106968. https://doi.org/10.1016/j.engappai.2023.106968.
(113) Adeyinka, A.; Oriola, O.; Tomomewo, O. S. An Innovative Approach for Oil Well Bottomhole Pressure Forecasting Using Kolmogorov-Arnold Neural Networks (KANs): A Case Study in an Offshore Oilfield. *Deep Resources Engineering* **2025**, 100233. https://doi.org/10.1016/j.deepre.2025.100233.

Appendices

Appendix A: Published Research Papers

This thesis is based on the following peer-reviewed journal publications:

Published Journal Articles

1. Nashed, S., Moghanloo, R., 2025. Replacing Gauges with Algorithms: Predicting Bottomhole Pressure in Hydraulic Fracturing Using Advanced Machine Learning. *Eng* 6, 73. https://doi.org/10.3390/eng6040073
2. Nashed, S., Moghanloo, R., 2025. Hybrid Symbolic Regression and Machine Learning Approaches for Modeling Gas Lift Well Performance. *Fluids* 10, 161. https://doi.org/10.3390/fluids10070161
3. Nashed, S., Moghanloo, R., 2025. Benchmarking ML Algorithms Against Traditional Correlations for Dynamic Monitoring of Bottomhole Pressure in Nitrogen-Lifted Wells. *Processes* 13, 2820. https://doi.org/10.3390/pr13092820