EVALUATION OF RESERVOIR INTER-WELL CONNECTIVITY USING MACHINE LEARNING TECHNIQUES: A CASE STUDY OF THE ALGERIAN MESDAR OIL FIELD

Fatima Kabli, Abdelhakim Benali, Hadjira Bekhadda, AND *Hadjer Benhmad**

I. Introduction

In petroleum production systems, crude oil is commonly extracted from reservoirs through fluid injection techniques. These operations are inherently

---

*Fatima Kabli received her Ph.D. in Computer Science from Tahar Moulay University of Saïda, Algeria, in 2018. She also holds a Master's degree in Knowledge Engineering in Computer Science (2013) and a Bachelor's degree in Computer Science from Oran 1 Ahmed Ben Bella University. She is currently an Assistant Professor at the National Polytechnic School of Oran, Maurice Audin (ENPO), Algeria. Her research interests include artificial intelligence, machine learning, and deep learning, with a particular emphasis on applications in industrial and healthcare domains. Her work has been published in international journals such as *IGI Global* and *Inderscience* and presented at conferences organized by the IEEE.

Abdelhakim Benali earned a degree in Petroleum Instrumentation from the Institute of Maintenance and Industrial Safety in Oran in 2009, followed by a Master's degree in Innovation of Industrial Systems and Products from the National Polytechnic School of Oran in 2012. He subsequently obtained a professional Master's degree in Reservoir Engineering from the Algerian Petroleum Institute in 2015 and is currently pursuing a Ph.D. in numerical fluid simulation and reservoir engineering. Since 2015, he has served as a Reservoir Engineer in Petroleum Engineering & Development at Sonatrach. He is also the founder of the technological platform AL-FARABI. His research and professional interests focus on integrated reservoir management and the adoption of emerging technologies to support the energy transition, with an emphasis on innovation in the oil and gas industry and contributions to climate change mitigation. (continued)

The Journal of Energy and Development, Vol. 50, Nos. 1-2

complex due to the interconnected dynamics between wells. Production wells share surface facilities for extraction and collection, while injection wells rely on common water-injection systems. At the subsurface level, reservoir structures further intensify the degree of interaction between production and injection wells, creating intricate connectivity patterns that must be managed effectively.

Over the past two decades, the study of well connectivity has become central to reservoir evaluation and development. Connectivity analysis not only supports more accurate reserve estimation but also facilitates strategic decision-making in well placement and long-term field management. By enhancing understanding of inter-well relationships, operators can obtain a more comprehensive view of reservoir heterogeneity. Such insights enable optimization of injection rates, improved planning for production wells, and significant reductions in drilling costs.

A variety of approaches have been developed to assess well connectivity, including engineering methods, reservoir simulations, and data-driven techniques. Traditional engineering tests—such as tracer, interference, and impulse tests—remain valuable tools but are often costly and time-intensive, which limits their scalability in large-scale petroleum applications. Numerical reservoir modeling offers another pathway, yet it too can be resource intensive.

In this context, Artificial Intelligence (AI) has emerged as a transformative alternative. By replicating aspects of human cognitive processes, AI offers powerful capabilities for analyzing complex, nonlinear data relationships at scale. Its application is expanding rapidly across multiple sectors, including healthcare (Nassif et al., 2022; Fatima et al., 2023), finance (Ahmed et al., 2022; Pallathadka et al., 2023), agriculture (Vyas et al., 2022; Poornappriya et al., 2022), transportation (Chu et al., 2022), education (Tang et al., 2022), robotics (Marcos-Pablos et al., 2022), and industry (Sundaram et al., 2014; Xu et al., 2022; Mathew and Brintha, 2023; Yan et al., 2020; Zhang et al., 2022; Fatima et al., 2024). Within the oil and gas sector, AI presents a promising frontier for advancing reservoir characterization, connectivity analysis, and production optimization.

Hadjira Bekhadda received both her Engineering degree and her Master's degree in Information Systems from the National Polytechnic School of Oran (ENPO), Algeria, in 2022. She is currently pursuing a Master's degree in Data Science and Engineering at Aix-Marseille University, France. Her research interests include artificial intelligence, machine learning, and deep learning. In addition to her academic studies, she has worked on projects related to data analysis and information retrieval, reflecting her interest in bridging theoretical research with practical applications.

Hadjer Benhmad received both her Engineering degree and her Master's degree in Information Systems from the National Polytechnic School of Oran (ENPO), Algeria, in 2022. Her research focuses on artificial intelligence, information systems, and management. She is currently employed as a Computer Engineer and is engaged in projects applying artificial intelligence to real-world challenges.

Industry 4.0 represents a new phase of industrial transformation, characterized by the integration of advanced digital technologies into production and operational systems. Within this paradigm, Artificial Intelligence (AI) plays a particularly critical role in the oil and gas sector, where it can be leveraged to optimize and automate production processes, strengthen safety protocols, and enhance risk management. AI-driven systems support real-time decision-making by identifying potential operational issues before they escalate into critical failures. Their integration into oil and gas operations has been shown to reduce lead times, enable more flexible and adaptive processes, and lower the incidence of workplace injuries and fatalities.

In recent years, machine learning (ML) has demonstrated considerable promise in the analysis of reservoir characteristics. Unlike traditional approaches, which rely on mathematical formulations grounded in physical laws to quantify inter-well connectivity, data-driven methods employ ML algorithms to detect and learn correlations within large datasets through iterative training. This enables accurate predictions of well connectivity without the need for explicit physical models. The practicality, efficiency, and scalability of these techniques have drawn increasing attention in petroleum engineering research and applications.

Building on this foundation, the present study proposes an intelligent system designed to automate aspects of the production process by analyzing well connectivity. The system employs state-of-the-art machine learning techniques and is trained on real-world data from the Mesdar oil field, located on the Saharan platform in Algeria.

2. Related Works

A considerable body of research has explored the application of artificial intelligence (AI) algorithms for identifying inter-well connectivity (the Appendix provides the nomenclature list). Early contributions by Heffer et al. (1997), Refunjol (1996), and Soeriawinata and Kelkar (1999) introduced models employing metrics such as the Spearman Correlation Coefficient (SCC). Other studies have examined inter-well connectivity through capacitance-resistance models (CRM), offering a framework for quantifying dynamic well interactions. Subsequent research has expanded these approaches by integrating advanced statistical and machine learning techniques. For instance, Pallathadka et al. (2023) proposed a hybrid methodology combining support vector machines with multiple linear regression to estimate production rates. Similarly, Dinh and Tiab (2013) developed a Modified Capacitance-Resistance Model (MCRM) tailored to immiscible gas injection scenarios, incorporating variables such as production rate, bottom-hole pressure, gas density, and average reservoir pressure. More recently, Artun (2016) applied a CRM framework in combination with artificial neural networks (ANNs) using a synthetic reservoir model. This approach involves analyzing the weights of trained

ANN connection links based on injection and production histories, as well as locations. A key advantage of ANNs lies in their flexibility: they are entirely data-driven and do not require pre-specified relationships among process variables. Instead, they infer complex nonlinear relationships directly from the training data, making them well-suited for modeling reservoir heterogeneity and dynamic well interactions.

Drawing on bottom-hole pressure (BHP) and historical production data, Sen (2021) examined two reservoir models: one consisting of two layers with high permeability between pairs of injector and producer wells, and another representing a fully homogeneous system. The study introduced two widely applied data-driven approaches for forecasting production in waterflooded environments: the capacitance-resistance model (CRM), implemented through non-linear regression, and a recurrent neural network (RNN), which calculates connectivity dynamically at each time step. Both methods demonstrated notable efficiency and were considerably faster than conventional reservoir simulators. BHP analysis has also been investigated by Dinh and Tiab (2013), who applied an analytical multivariate linear regression framework to account for fluctuations in injectors and producers in waterflooded systems, ultimately deriving inter-well connectivity coefficients. In contrast, Liu et al. (2019) employed an artificial neural network (ANN) trained exclusively on BHP data—rather than historical injection and production records—to model well interactions in two synthetic cases: a homogeneous reservoir and an anisotropic reservoir.

Du et al. (2020) developed a deep learning (DL) framework that integrates three-dimensional convolutional neural networks (3D-CNN) with backpropagation (BP) networks to identify inter-well connectivity without relying on explicit physical models. The approach utilizes geological permeability models in combination with dynamic production data—including oil production rates, water cut, and injection pressure—generated through numerical simulations. The results indicate that this method achieves a closer correspondence to actual reservoir connectivity when compared with traditional modeling techniques.

Liu (2020) applied convolutional neural networks (CNN) and backpropagation (BP) techniques to analyze reservoirs containing impermeable interlayers that obstruct fluid flow. Using simulation technologies, experiments were conducted under varying conditions of permeability, interlayer angle, and injection pressure, with dynamic production data serving as input. The results showed that the predictions of the proposed model were consistent with those of traditional methods, while offering significantly reduced computation times.

In related work, Cheng et al. (2020) introduced a deep learning framework based on long short-term memory (LSTM) neural networks combined with global sensitivity analysis (GSA) to model the mapping relationships between production wells and surrounding injection wells in a synthetic reservoir system. The model utilized historical injection and production fluctuation data, while the extended Fourier amplitude sensitivity test (EFAST) was integrated with GSA to estimate connectivity indices. This methodology proved to be an efficient and robust approach for evaluating inter-well connectivity.

Cheng et al. (2019) employed long short-term memory (LSTM) networks in conjunction with artificial neural networks (ANNs) to improve the accuracy of liquid production forecasting based on historical well data. A sensitivity analysis was then conducted on the trained three-layer ANN model to simulate connectivity relationships between production wells and surrounding injectors.

Whereas, Demiryurek et al. (2008) applied neural networks to a comprehensive dataset of real production and injection records, constructing a sensitivity matrix to evaluate the influence of each injector on oil output. Their findings demonstrated that this approach effectively identified both the least and most impactful injectors, while also achieving higher correlation accuracy than the estimations provided by field engineers.

Yuan et al. (2021) proposed a particle swarm optimization–based system (PSOC4IC) for analyzing inter-well connectivity by integrating a denoising autoencoder capable of handling high-dimensional noisy data. The input dataset included parameters such as porosity, oil saturation, recoverable oil saturation, perforation state, and logging data. In comparison to conventional neural network approaches, which often suffer from reduced accuracy due to limited data availability, the PSOC4IC framework demonstrated superior performance in characterizing inter-well connectivity.

Huang et al. (2021) developed a machine learning–based approach for quantitatively analyzing the key control factors influencing oil saturation variation. By leveraging both static and dynamic field data, the method provides reservoir engineers with improved insights to support more informed decision-making.

In our paper, we present an alternative approach to rapidly predict inter-well connectivity using machine learning methods and model-agnostic interpretation techniques. The proposed methodology is divided into the following steps: First, based on dynamic data such as production data, injection rates, pressure, and static data, machine learning models are established to predict oil production rates, and their accuracy is compared to determine the best-performing model. Second, we analyze the impact of each parameter on oil production using machine learning techniques. Finally, we utilize the trained models to characterize the inter-well connectivity by applying model-agnostic interpretation methods.

3. Materials and Methods

3.1 General Architecture of Our Approach: In this study, machine learning methods are employed to develop a system for processing oil well data within a real reservoir model. The dataset, provided by the PED division of Sonatrach in Algeria, includes both geological and dynamic reservoir information. The primary objective of the system is to characterize inter-well connectivity, with a specific focus on identifying the factors that define the interactions between injection and production wells. The workflow is structured in several phases. First, data preprocessing is performed, including cleaning and outlier treatment. Second, multiple machine learning models are applied to predict oil production rates, with their performance evaluated and compared through appropriate scoring metrics. Third, the best-performing model is used to quantitatively assess the key controlling factors influencing oil production. Finally, inter-well connectivity is inferred from the trained regression model using the permutation feature importance method. The major phases are presented in Figure 1.

3.2 Data Collection and Preprocessing: Data collection and preprocessing represent a critical stage in the development of machine learning models, as they ensure the extraction of meaningful insights from complex variable relationships. The primary objective in this phase is to gather comprehensive information about the wells—both numerical and categorical—through collaboration with reservoir engineers, geologists, petrophysicists, and other domain specialists.

The dataset used in this study is derived from real reservoir data and includes information from both producing and injecting wells. It encompasses oil, water, and gas measurements from the Mesdar field, located in the central-northern Algerian Saharan platform, approximately 100 km southeast of the Hassi Messaoud giant oil field. The Mesdar field is part of the Rhourde el Baguel–Mesdar structural entity and covers an area of about 40 km^2. The oil-water contact lies at a depth of 3390 m below the surface. Reservoir fluids are characterized by low-viscosity oil (0.4 cP) with density values ranging from 0.604 g/cm^3 to 0.661 g/cm^3, and the field exhibits an average water saturation of 20.9%.

Figure 1
GENERAL PROCESS OF OUR WORK

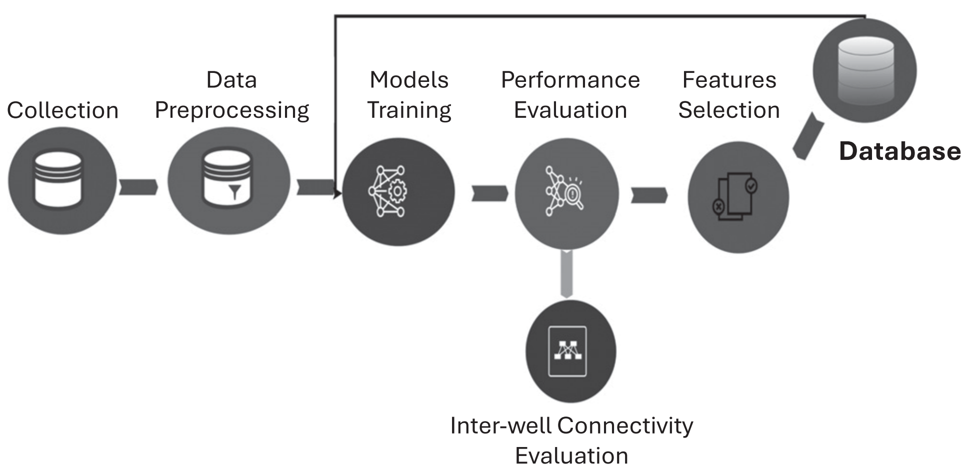


Table 1
CHARACTERISTICS OF RESERVOIR MODEL

Model Description	**Values**
Number of grid blocks	13351400
Grid Block size	$277 \times 482 \times 100$
Reservoir Permeability	11(Md)
Reservoir porosity	5,37

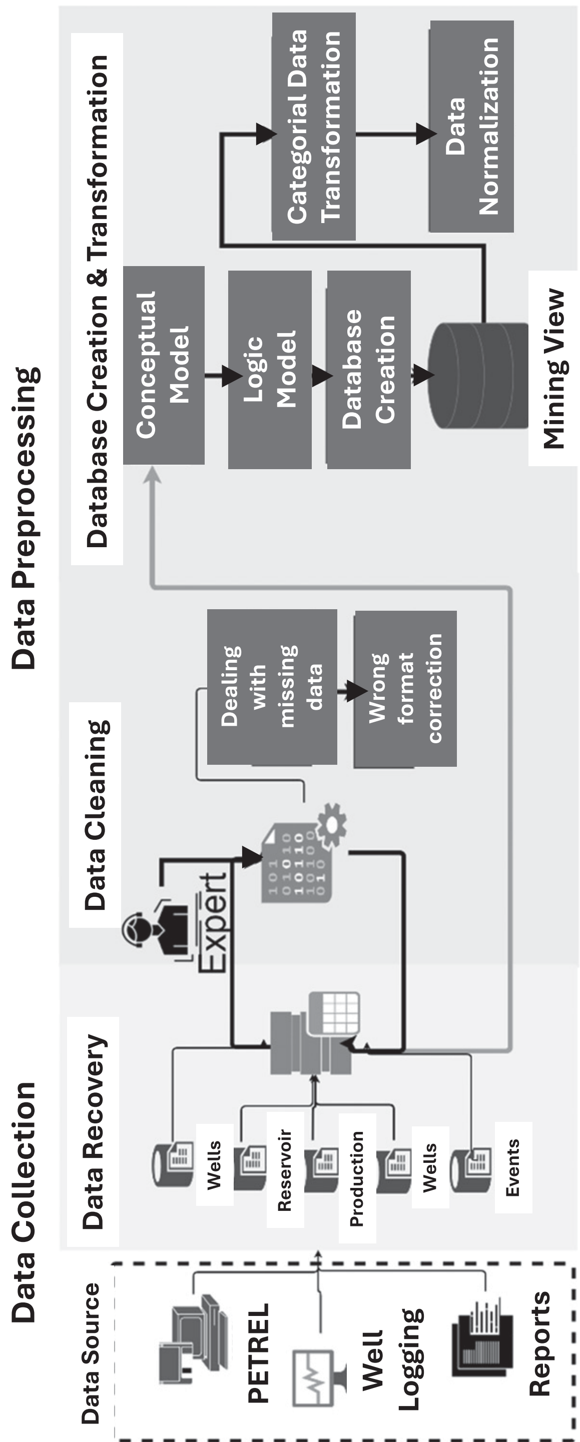
The dataset encompasses a wide range of static and dynamic reservoir information. Static data include depth, geological characteristics, and well logs (petrophysical properties) such as permeability, porosity, oil saturation, and the net-to-gross (NTG) ratio. Dynamic data comprise oil production, water injection, and pressure records obtained through routine daily and monthly monitoring.

The detailed dataset is derived from multiple sources, including seismic interpretation, logging reports, core analyses, and reservoir studies conducted on 12 wells (nine producers and three injectors). These wells have a production history spanning from February 1969 to 2016.

As illustrated in Figure 2, data preprocessing and cleaning are essential prior to applying machine learning algorithms, given the inherent noise, unreliability, and incompleteness of real-world data. Without proper preparation, such issues can lead to inaccurate results. Preprocessing enhances the quality of the dataset and significantly improves model performance. This process in our workflow involves three main steps:

- *Data Cleaning:* Ensuring data quality and consistency by detecting and correcting errors, addressing anomalies, and handling missing values through either replacement or removal.
- *Data Integration:* Constructing a unified dataset suitable for model training. In data science, most machine learning algorithms require input in the form of a single matrix, commonly referred to as the "Mining View." However, raw data collected from field operations rarely adhere to this structure. To resolve this, we employed the PostgreSQL database management system to build a coherent and structured database.
- *Data Transformation:* Converting categorical features into numerical representations, since most algorithms cannot directly process categorical inputs. For this purpose, we applied the widely used "OneHotEncoder" technique. Additionally, data sampling and transformation operations—such as calculating aggregates (e.g., sums or averages)—were performed. These iterative operations refine the dataset, enabling clearer problem understanding and more effective control.

Figure 2
GENERAL ARCHITECTURE OF APPROACH



3.3 Prediction of Oil Production Rate: Following data collection and preprocessing, we developed an approach to train multiple regression models for estimating oil production (Figure 3). The models were designed to reproduce historical production rates by incorporating static reservoir characteristics and injection rates as input variables. By evaluating different combinations of datasets and algorithms, the framework allows for improved understanding of the interactions between injection and production wells.

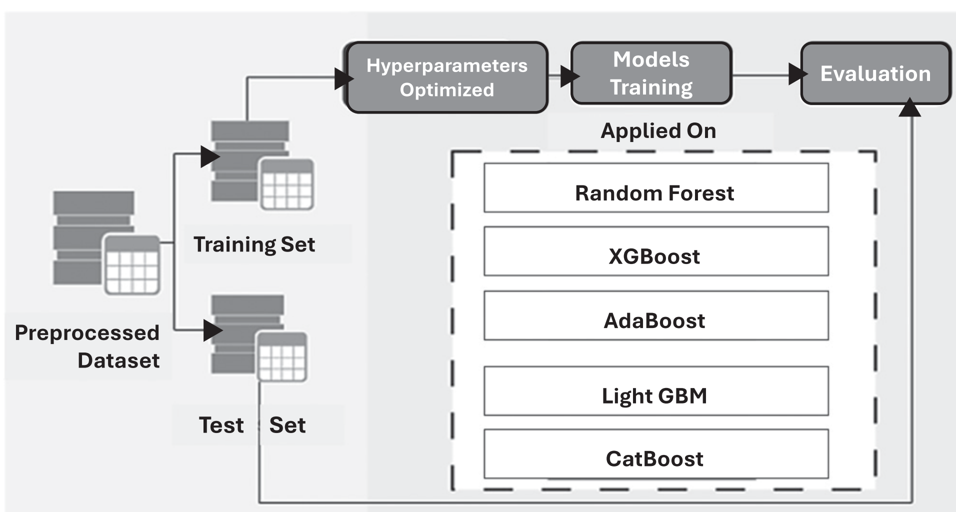
The dataset was divided into training and testing subsets, enabling a systematic comparison of model performance across different algorithms. Specifically, the first subset covers the production period from 1969 to 2006, while the second subset spans the years 2007 to 2016.

3.3.1 Hyperparameter Optimization: In machine learning, hyperparameters are predefined settings that govern the internal behavior of an algorithm and must be specified prior to the learning process. Unlike model parameters, which are randomly initialized and iteratively optimized during training, hyperparameters remain fixed throughout a training run and strongly influence model performance. Common examples include the number of trees, learning rate, and maximum tree depth.

The process of hyperparameter tuning generally involves the following steps:

1. Selecting candidate hyperparameter values, either intuitively or at random.
2. Building a model based on the selected hyperparameters.
3. Training the model and recording its performance.
4. Repeating the process with different hyperparameter values.

Figure 3
OIL PRODUCTION PREDICTION MODELS



The optimal configuration is then chosen based on the best-performing model. In this study, we employed the Grid Search method to systematically identify the most suitable hyperparameters for each algorithm, thereby enhancing efficiency and predictive accuracy.

3.3.2 Machine Learning Algorithms: Using the preprocessed database, we selected five machine learning algorithms to develop regression models aimed at uncovering hidden relationships between the dependent variable (oil production) and various independent variables. The selected methods include ensemble approaches, such as Random Forest based on the bagging technique, as well as algorithms that employ the boosting technique (Alban, 2017).

Random Forest is a supervised ensemble learning algorithm used for both classification and regression, originally proposed by Breiman (2001). It constructs an ensemble of NNN base learners, typically decision trees, each trained on a randomly selected subset of the data and features. By introducing randomness in both sampling and feature selection, the algorithm reduces variance and mitigates overfitting. For prediction, the outputs of all trees are aggregated to produce a more accurate and robust result. In regression tasks, the final prediction is obtained by averaging the outputs of the individual trees, whereas in classification tasks it is determined by the majority vote among them (Freund et al., 1996; Zocca et al., 2017).

Adaptive Boosting (AdaBoost, ADBT) is an iterative ensemble algorithm that trains multiple weak classifiers on the same dataset and then combines them to form a stronger final model (Chen et al., 2016). The method improves prediction accuracy and exhibits robustness against overfitting. XGBoost is a scalable, supervised learning framework for gradient boosting decision trees. It operates sequentially, combining the predictions of many weak learners to achieve higher accuracy. Its self-improving mechanism relies on optimizing a loss function, with performance controlled by a large number of tunable hyperparameters (Wang et al., 2019).

LightGBM (LGBM), or Light Gradient Boosting Machine, is an advanced implementation of gradient boosting that introduces gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB). These innovations improve computational efficiency, reduce memory usage, and enhance performance, making LGBM particularly well suited for large-scale datasets (Tang et al., 2020).

CatBoost (Categorical Boosting) is specifically designed to handle categorical variables efficiently by converting them into numerical features using statistical transformations derived from feature combinations. This approach improves model accuracy and reduces the need for extensive preprocessing of categorical data (Prokhorenkova et al., 2017).

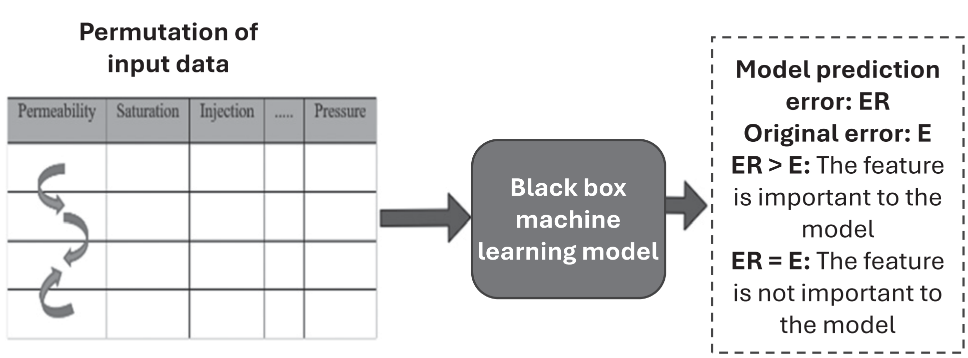
3.3.3 Quantitative Analysis of the Main Controlling Factors of Oil Production/ Feature Selection: During reservoir development, oil production rates are influenced by a wide range of static and dynamic parameters. The initial dataset contained 42 input features describing well characteristics, including production timelines, pressure data, additional production indicators, water injection information, and well log measurements. However, the inclusion of irrelevant or redundant attributes can negatively affect the accuracy and efficiency of machine learning algorithms. To address this, we performed a systematic feature selection process aimed at identifying the factors most strongly associated with oil production.

To quantify the contribution of each variable to production variability, we applied machine learning–based quantitative analysis using the Permutation Feature Importance method. This approach evaluates the relevance of individual features by measuring the increase in prediction error when a given feature's values are randomly permuted, while keeping all other variables unchanged. Features that, when permuted, cause a significant increase in error are deemed important, whereas those that produce little or no change are considered less relevant or redundant.

The optimal regression models trained in this study were used as the basis for these analyses. By applying permutation tests, we calculated the impact factor of each variable on oil production, thereby refining the dataset, reducing dimensionality, and ultimately improving model performance. Figure 4 illustrates the detailed process of calculating feature importance using this method.

3.3.4 Evaluation of Inter-Well Connectivity: Inter-well connectivity was evaluated by analyzing the relationship between injection data and oil production rates. Initially, the Pearson correlation coefficient was used to assess linear correlations and identify the potential influence of each injector well on the corresponding producer wells. However, while machine learning models are effective for prediction, they are often regarded as black-box approaches, offering limited interpretability regarding the relationships between input features and outputs. To enhance interpretability, we employed the Permutation Feature Importance method. This technique involves permuting individual input features (injection rates) in the trained model and measuring the resulting increase in root mean square error (RMSE)

Figure 4
FEATURE SELECTION BASED ON PERMUTATION DATA



relative to the baseline. A greater increase in error indicates a stronger contribution of that feature to model accuracy. In addition, we applied feature importance methods based on reduction criteria such as Gini impurity and entropy, which assign relative importance scores to input variables. Together, these approaches enabled us to identify the most and least influential injector wells, thereby providing a more accurate understanding of inter-well connectivity.

4. Experimental Results

Following data collection and preprocessing, a substantial volume of valuable reservoir data was compiled. To effectively evaluate and validate our proposed approach, an integrated dataset was constructed and used to train multiple machine learning models. For this purpose, a dedicated database—referred to as the "Reservoir Database"—was developed. Table 2 summarizes the number of instances included in this dataset.

To evaluate the effectiveness of the proposed approaches, we present the experimental results in terms of established performance metrics in Table 3. For production rate prediction, a regression-based framework was implemented using five machine learning models trained on the constructed dataset. The analysis pursued two main objectives: (i) to accurately predict production rates and (ii) to reduce dimensionality by retaining only the most influential features. Model performance was assessed using the Root Mean Square Error (RMSE) as the primary error metric and accuracy scores for overall evaluation. Furthermore, hyperparameter optimization was conducted via grid search across the different algorithms to ensure the selection of optimal model configurations.

The exhaustive hyperparameter search identified the optimal configuration for the Random Forest regression model as follows:

- Number of trees in the forest (n_estimators): 15
- Maximum number of features considered for node splitting (max_features): "auto"

Table 2
NUMBER OF INSTANCES FOR EACH TABLE
IN THE CONSTRUCTED DATABASE

Table	Numbers of instance
Well	12
Reservoir	24
Production	5184
Injection	1728
Events	5736
Final table	61920*41

Table 3
ASSESSMENT OF PRODUCTION RATE PREDICTION MODELS

Model	Accuracy	MAE	MSE	RMSE
Random Forest	0.965	208.317	652473.251	807.758
XGBoost	0.9710	251.510	543757.731	737.399
CatBoost	0.968	310.898	583375.997	763.790
LightGBM	0.960	351.102	740333.198	860.4261
AdaBoost	0.8573	867.232	2674567.706	1635.410
Linear regression	0.8494	767.882	2495874.368	988.417
Polynomial regression	0.7442	1087.108	1074268.706	4522.554
Decision tree	0.8943	625.531	2142258.706	9854.110
SVM	0.7271	1224.430	1474567.706	6997.487
Lasso	0.7793	977.741	2922148.706	10974.740

- Maximum depth of all decision trees (max_depth): 25
- Minimum number of samples required to split a node (min_samples_split): 5
- Minimum number of data points allowed in a leaf node (min_samples_leaf): 1

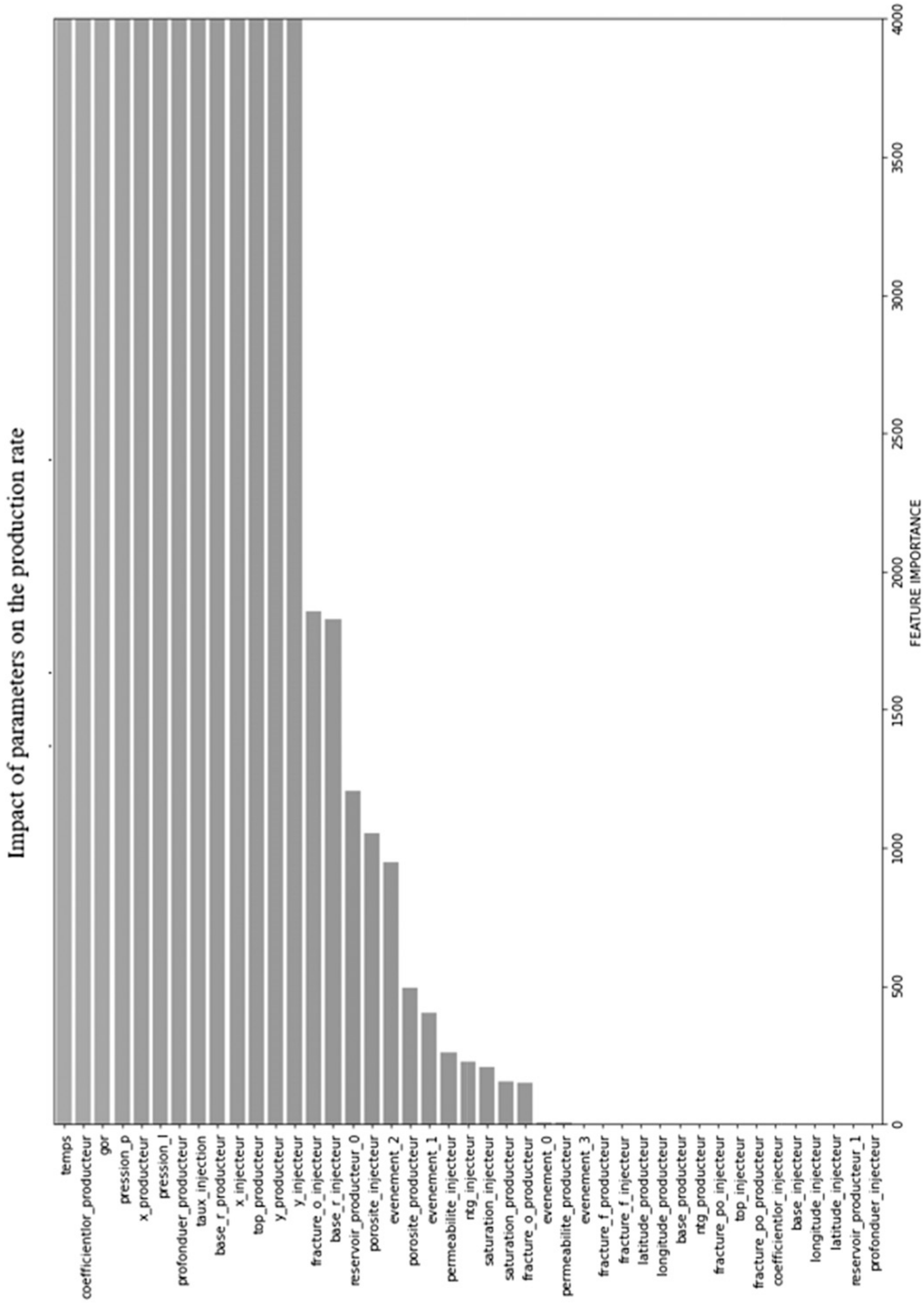
This parameter combination yielded an accuracy score of 0.96.

As shown in Table 3, nonlinear regression models—particularly XGBoost and Random Forest—clearly outperformed the other algorithms. Both achieved high determination coefficients (R2) of 0.971 and 0.965, respectively, along with low RMSE values of 737.399 and 807.758. The smaller error and R2 values close to 1 confirm the superior predictive performance of these models. In contrast, AdaBoost exhibited the weakest performance, with an accuracy of 0.857 and an RMSE of 1635.410, while the remaining models achieved only modest results. Overall, these findings are highly encouraging given the constructed dataset.

The distinction between these models lies in their underlying mechanisms. XGBoost employs a boosting process, iteratively improving prediction accuracy by assigning greater weight to difficult-to-predict values during training, thereby forcing the model to improve progressively. Random Forest, on the other hand, aggregates predictions through majority voting across independent decision trees, whereas XGBoost combines predictions sequentially throughout the learning process.

To further interpret model behavior, the XGBoost model was applied to quantify the impact of each feature on oil production using the Permutation Feature Importance method. The results are illustrated in Figure 5. The accompanying bar chart provides a visual ranking of feature importance: the x-axis represents the importance score, where longer bars denote greater influence, while the y-axis lists the features in descending order of significance, from most to least important.

Figure 5
RESULTS OF THE IMPACT OF PARAMETERS ON PRODUCTION HISTORY



Oil production is shown to be primarily influenced by operating time, reflecting the underlying reservoir development dynamics. The Gas-Oil Ratio (GOR) and the Lorentz coefficient follow as the second and third most significant factors, respectively. Other highly relevant features include producer pressure, injector pressure, and injection rate, emphasizing the critical role of injection–production well models in enhancing oil recovery and exploiting remaining reserves.

This feature selection process enabled us to retain only the most informative variables while eliminating redundant or less significant ones. By reducing the dimensionality of the dataset, computational efficiency was improved and model performance was optimized. Table 4 summarizes the prediction results for oil production rates using the refined dataset. The findings indicate that the AdaBoost model consistently produced the lowest scores, while the performance of other models was moderate. Among the constructed models (Random Forest, XGBoost, LightGBM, and AdaBoost), substantial improvements were observed after incorporating feature importance analysis compared to results obtained with the original dataset. For CatBoost, performance metrics improved slightly; however, the difference between results on the original and reduced datasets was minimal. Overall, the comparative evaluation demonstrates that XGBoost and Random Forest achieved the highest predictive accuracy, making them the most effective models for assessing inter-well connectivity.

The Pearson correlation coefficient was applied to evaluate the degree of association between water injection rates and oil production responses. Based on the correlation matrix, inter-well connectivity can be inferred. The results, illustrated in Figure 6, show that injector IG22 exhibits strong positive correlations with producers P1, P3, P4, P5, and P6, while all injectors demonstrate negative correlations with producers P8, P9, and P20. The correlation between water injection and water production further supports evidence of connectivity. Given the limited number of injectors (three in total), their correlations with producers are generally robust, with values that are nearly identical across wells.

Reservoir connectivity was also assessed using the Random Forest model in conjunction with the Permutation Feature Importance method. In this framework,

Table 4
EVALUATION OF OIL PRODUCTION PREDICTION MODIFIED MODELS

E^2	Accuracy	MAE	MSE	RMSE
Random Forest	0.9703	225.20	555184.447	745.10
XGBoost	0.97100	251.50	543752.258	737.39
CatBoost	0.968	311.73	588465.682	767.11
LightGBM	0.965	347.43	720548.152	848.85
AdaBoost	0.863	730.32	2565385.996	1601.6

Figure 6
CORRELATION MATRIX OF INJECTION RATES AND LIQUID PRODUCTION RATES
(The table displays the increase in RMSE (Root Mean Square Error) of the Random Forest model when the injection rate is permuted)

	P1	P2	P3	P4	P5	P6	P8	P9	P20	IG20	IG21	IG22
P1I	1	-0.09	0.39	0.25	0.30	0.31	0.54	0.67	0.30	0.06	0.09	0.13
P2I	-0.09	1	-0.06	0.08	-0.04	0.04	-0.07	0.05	0.14	0.15	0.08	0.09
P3I	0.39	-0.06	1	0.67	0.73	0.65	0.15	0.32	-0.32	0.39	0.36	0.47
P4I	0.25	0.08	0.67	1	0.83	0.78	-0.01	0.09	-0.31	0.50	0.51	0.58
P5I	0.30	-0.04	0.73	0.83	1	0.80	0.06	0.18	-0.31	0.33	0.40	0.57
P6I	0.31	0.04	0.65	0.78	0.80	1	0.08	0.17	-0.31	0.34	0.40	0.46
P8I	0.54	-0.07	0.15	-0.01	0.06	0.08	1	0.74	0.40	-0.16	-0.13	-0.14
P9I	0.67	0.05	0.32	0.09	0.18	0.17	0.74	1	0.37	-0.02	-0.06	-0.03
P20I	0.30	0.14	-0.32	-0.31	-0.31	-0.31	0.40	0.37	1	-0.20	-0.21	-0.24
IG20I	0.06	0.15	0.39	0.50	0.33	0.34	-0.16	-0.02	-0.20	1	0.79	0.75
IG21I	0.09	0.08	0.36	0.51	0.40	0.40	-0.13	-0.06	-0.21	0.79	1	0.67
IG22I	0.13	0.09	0.47	0.58	0.57	0.46	-0.14	-0.03	-0.24	0.75	0.67	1

an injector is considered influential when permuting its injection rate results in a measurable increase in prediction error. Conversely, if permuting the injection rate does not significantly alter the error, the injector is deemed to have little or no impact on production.

Based on the results presented in tables 5 and 6, the application of both techniques for evaluating inter-well connectivity with the trained Random Forest model produced largely consistent outcomes. Injector IG22 demonstrated the strongest correlations with the majority of producers, ranking highest in influence over P9, P4, P5, P8, P1, P2, and P6, in that order. Injector IG21 ranked second, exhibiting similar relationships with the same group of producers. Notably, producer P20 showed no detectable connectivity with any of the injectors.

Inter-well connectivity was further assessed using the trained XGBoost model, applying both Permutation Feature Importance and Feature Importance methods. The corresponding Table 7 reports the increase in Root Mean Square Error (RMSE) observed when injection rate variables were permuted, thereby quantifying their relative impact on production predictions.

From Tables 7 and 8, it can be observed that the evaluation of inter-well connectivity using the Permutation Feature Importance method produced results that were slightly different from those obtained with the second technique applied to the trained XGBoost model. In both cases, the strongest connections consistently involved producers P1, P2, P4, P5, and P9. However, for the remaining

Table 5
INTER-WELL CONNECTIVITY USING THE PERMUTATION IMPORTANCE METHOD OF THE RANDOM FOREST MODEL

	Injector		
Producer	**IG20**	**IG21**	**IG22**
P1	52715.05605	209010.67785	210270.02932
P2	36272.20799	113974.65223	142951.38256
P3	15307.69143	18589.38938	42435.96893
P4	165405.42844	312734.97538	2195668.55170
P5	255338.59236	383466.28174	1901431.86188
P6	72008.00258	43035.32299	55053.26799
P8	22893.02469	59046.43151	327563.70840
P9	708282.95416	366151.43782	2469778.34832
P20	0	0	0

producer–injector pairs, the relative importance rankings were interchanged. Notably, producer P20 showed no evidence of connectivity with any injector across both methods.

Overall, the evaluation of inter-well connectivity using two model-interpretation techniques—applied to both the Random Forest and XGBoost regressors—demonstrated that these approaches treat the models as black boxes and are theoretically applicable to either algorithm. With the exception of the correlation coefficient method, all techniques successfully identified the most relevant injector

Table 6
INTER-WELL CONNECTIVITY USING THE FEATURE IMPORTANCE METHOD OF THE RANDOM FOREST MODEL

	Injector		
Producer	**IG20**	**IG21**	**IG22**
P1	1.13081462e-02	2.88731543e-02	3.00479628e-02
P2	9.3219876e-03	1.5436468e-02	2.7998688e-02
P3	7.20996004e-03	1.70119370e-02	2.09796469e-02
P4	1.93882895e-03	9.36012165e-03	3.36329127e-02
P5	2.18202860e-03	5.97647037e-03	1.72699671e-02
P6	1.56354944e-02	1.01161908e-02	8.49555978e-03
P8	1.29807507e-03	2.53715939e-03	7.12704558e-03
P9	6.29659234e-03	5.03177193e-03	1.26875571e-02
P20	0	0	0

Table 7
INTER-WELL CONNECTIVITY USING THE PERMUTATION IMPORTANCE METHOD OF XGBOOST

	Injector		
Producer	**IG20**	**IG21**	**IG22**
P1	27878.30246	75974.95888	119273.02074
P2	21066.49508	98688.27160	105778.09486
P3	22815.17875	8326.06557	75188.29346
P4	20942.60674	96680.47542	116224.93783
P5	968697.49393	383712.93496	2904178.74336
P6	77058.65621	79050.05321	112969.49646
P8	48138.29100	44804.53431	414730.09066
P9	1088349.52111	227695.13029	2153262.15262
P20	0	0	0

Table 8
INTER-WELL CONNECTIVITY USING THE FEATURE IMPORTANCE METHOD OF XGBOOST MODEL

	Injector		
Producer	**IG20**	**IG21**	**IG22**
P1	4.0351469e-03	7.6468843e-03	7.8535862e-03
P2	9.3219876e-03	1.5436468e-02	2.7998688e-02
P3	3.2024083e-03	5.8354880e-03	6.8069594e-03
P4	9.3219876e-03	1.5436468e-02	2.7998688e-02
P5	3.2323026e-03	2.2292766e-03	1.0074890e-02
P6	6.8719843e-03	5.0168661e-03	4.4048126e-03
P8	1.5776948e-03	1.8159067e-03	4.1287607e-03
P9	1.9899094e-03	2.3411305e-03	3.2328381e-03
P20	0	0	0

for producer P2. Based on these findings, we conclude that the Feature Importance method provides superior results. Its primary advantage lies in its flexibility, as it can be implemented effectively without the need for specialized domain expertise.

5. Conclusion and Perspectives

This study presented a machine learning–based framework for evaluating inter-well connectivity using both static and dynamic reservoir data. The workflow began with the collection and preprocessing of well data, followed by the application of ensemble learning algorithms—boosting and bagging—to train predictive models of oil production. Model-agnostic interpretation techniques were then employed to analyze the contribution of input variables, enabling the selection of the most relevant features. Using these refined features, an optimized predictive model of oil production was developed. Finally, inter-well connectivity coefficients were derived from the trained models and compared with results obtained through a conventional static method, allowing for cross-validation of outcomes. The analysis and discussion of results lead to the following conclusions:

- Feature impact analysis effectively identifies the most influential parameters while eliminating redundant or irrelevant variables, thereby improving predictive accuracy and model robustness.
- Among the predictive models tested, Random Forest and XGBoost achieved the highest accuracy in forecasting oil production rates, both exceeding 97%.
- Model-agnostic interpretation methods proved to be efficient tools for evaluating inter-well connectivity, offering interpretable insights into injector–producer relationships.

As a future direction, the proposed methodology will be applied to a real industrial oil field to rigorously evaluate the performance of the hybrid algorithm under operational conditions. This validation will account for complex variations such as producer-to-injector conversions and the deliberate suspension of production and injection processes. Incorporating these dynamic operational scenarios will enhance the practical relevance of the approach and demonstrate its robustness in addressing the intricate challenges of real-world reservoir management.

REFERENCES

Ahmed, S., M. M. Alshater, A. El Ammari, and H. Hammami. 2022. "Artificial Intelligence and Machine Learning in Finance: A Bibliometric Review." *Research in International Business and Finance* 61: 101646. https://doi.org/10.1016/j.ribaf.2022.101646.

Alban, T. 2017. "Algorithmes de Boosting – AdaBoost, Gradient Boosting, XGBoost." *Datascientest*. Accessed January 2, 2022. Algorithmes de Boosting – AdaBoost, Gradient Boosting, XGBoost.

Artun, E. 2016. "Characterizing Reservoir Connectivity and Forecasting Waterflood Performance Using Data-Driven and Reduced-Physics Models." In *SPE Western Regional Meeting*, SPE-180488. SPE. https://doi.org/10.2118/180488-MS.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45: 5–32. https://doi.org/10.1023/A:1010933404324.

Chen, T., and C. Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In Proceedings of the *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. https://doi.org/10.1145/2939672.2939785.

Cheng, H., X. Han, P. Zeng, H. Yu, E. Osipov, and V. Vyatkin. 2019. "ANN Based Interwell Connectivity Analysis in Cyber-Physical Petroleum Systems." In *2019 IEEE 17th International*

Conference on Industrial Informatics (INDIN), 199–205. IEEE. https://doi.org/10.1109/INDIN41052.2019.8972285.

Cheng, H., V. Vyatkin, E. Osipov, P. Zeng, and H. Yu. 2020. "LSTM Based EFAST Global Sensitivity Analysis for Interwell Connectivity Evaluation Using Injection and Production Fluctuation Data." *IEEE*. https://doi.org/10.1109/ACCESS.2020.2985230.

Chu, S. T., G. J. Hwang, and Y. F. Tu. 2022. "Artificial Intelligence-Based Robots in Education: A Systematic Review of Selected SSCI Publications." *Computers and Education: Artificial Intelligence* 3: 100091. https://doi.org/10.1016/j.caeai.2022.100091.

Demiryurek, U., F. Banaei-Kashani, C. Shahabi, and F. G. Wilkinson. 2008. "Neural-Network Based Sensitivity Analysis for Injector-Producer Relationship Identification." In *Intelligent Energy Conference and Exhibition.* OnePetro. https://doi.org/10.2118/112124-MS.

Dinh, A. V., and D. Tiab. 2013. "Inferring Interwell Connectivity in a Reservoir from Bottomhole Pressure Fluctuations in Hydraulically Fractured Vertical Wells, Horizontal Wells, and Mixed Wellbore Conditions." In *SPE Oklahoma City Oil and Gas Symposium/Production and Operations Symposium*, SPE-164482. SPE. https://doi.org/10.2118/164482-MS.

Du, S., R. Wang, C. Wei, Y. Wang, Y. Zhou, J. Wang, and H Song. 2020. "The Connectivity Evaluation Among Wells in Reservoir Utilizing Machine Learning Methods." *IEEE Access* 8: 47209–19. https://doi.org/10.1109/ACCESS.2020.2979358.

Fatima, K., W. M. A. Walid, B. Ouahiba, and A. Mohammed. 2023. "Semantic Segmentation of CT Images for the Detection of Coronavirus Lesions: The Case of Algeria." In *2023 International Conference on Advances in Electronics, Control and Communication Systems* (ICAECCS), 1–6. IEEE. http://dx.doi.org/10.1109/ICAECCS56710.2023.10105011.

Fatima, K., S. Abdeldjalil, and M. Mohamed. 2024. "Detection of Weld Defects in Patch-Based Radiographic Images Using Deep Learning." In *2024 8th International Conference on Image and Signal Processing and Their Applications* (ISPA), 1–10. IEEE. http://dx.doi.org/10.1109/ISPA59904.2024.10536751.

Freund, Y., R. E. Schapire, et al. 1996. "Experiments with a New Boosting Algorithm." In *Proceedings of the 13th International Conference on Machine Learning*, 148–56. Citeseer.

Heffer, K. J., R. J. Fox, C. A. McGill, and N. C. Koutsabeloulis. 1997. "Novel Techniques Show Links Between Reservoir Flow Directionality, Earth Stress, Fault Structure and Geomechanical Changes in Mature Waterfloods." *SPE Journal* 2 (2): 91–98. https://doi.org/10.2118/30711-PA.

Huang, R., C. Wei, J. Yang, X. Xu, B. Li, S. Wu, and L. Xiong. 2021. "Quantitative Analysis of the Main Controlling Factors of Oil Saturation Variation." *Geofluids* 2021: 1–12. http://dx.doi.org/10.1155/2021/6515846.

Liu, J. 2020. "Potential for Evaluation of Interwell Connectivity Under the Effect of Intraformational Bed in Reservoirs Utilizing Machine Learning Methods." *Geofluids* 2020: 1–13. https://doi.org/10.1155/2020/1651549.

Liu, W., W. D. Liu, and J. Gu. 2019. "Reservoir Inter-Well Connectivity Analysis Based on a Data Driven Method." In *Abu Dhabi International Petroleum Exhibition & Conference.* OnePetro.

Marcos-Pablos, S., and F. J. García-Peñalvo. 2022. "Emotional Intelligence in Robotics: A Scoping Review." In *New Trends in Disruptive Technologies, Tech Ethics and Artificial Intelligence: The DITTET Collection* 1, 66–75. Cham: Springer International Publishing. http://dx.doi.org/10.1007/978-3-030-87687-6_7.

Mathew, D., and N. C. Brintha. 2023. "Artificial Intelligence in the Field of Textile Industry: A Systematic Study on Machine Learning and Neural Network Approaches." In *Recent Trends in Computational Intelligence and Its Application: Proceedings of the 1st International Conference on Recent Trends in Information Technology and Its Application* (ICRTITA 22), 222. CRC Press. https://doi.org/10.1201/9781003388913.

Nassif, A. B., M. A. Talib, Q. Nasir, Y. Afadar, and O. Elgendy. 2022. “Breast Cancer Detection Using Artificial Intelligence Techniques: A Systematic Literature Review.” *Artificial Intelligence in Medicine* 127: 102276. https://doi.org/10.1016/j.artmed.2022.102276.

Pallathadka, H., E. H. Ramirez-Asis, T. P. Loli-Poma, K. Kaliyaperumal, R. J. M. Ventayen, and M. Naved. 2023. “Applications of Artificial Intelligence in Business Management, E-Commerce and Finance.” *Materials Today: Proceedings* 80 (6): 2610–13. https://doi.org/10.1016/j.matpr.2021.06.419.

Poornappriya, T. S., and R. Gopinath. 2022. “Rice Plant Disease Identification Using Artificial Intelligence Approaches.” *International Journal of Electrical Engineering and Technology* 11 (10): 392–402. http://dx.doi.org/10.34218/IJEET.11.10.2020.050.

Prokhorenkova, L., G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin. 2017. “CatBoost: Unbiased Boosting with Categorical Features.” *arXiv preprint arXiv:1706.09516*. https://doi.org/10.48550/arXiv.1706.09516.

Refunjol, B. T. 1996. “Reservoir Characterization of North Buck Draw Field Based on Tracer Response and Production/Injection Analysis.” PhD diss., Colorado School of Mines.

Sen, D. 2021. “Capacitance Resistance Model and Recurrent Neural Network for Well Connectivity Estimation: A Comparison Study.” *arXiv preprint arXiv:2109.08779*. https://doi.org/10.48550/arXiv.2109.08779.

Soeriawinata, T., and M. Kelkar. 1999. “Reservoir Management Using Production Data.” In *SPE Mid-Continent Operations Symposium*. OnePetro. https://doi.org/10.2118/52224-MS.

Sundaram, M., J. P. Jose, and G. Jaffino. 2014. “Welding Defects Extraction for Radiographic Images Using C-Means Segmentation Method.” In *2014 International Conference on Communication and Network Technologies*, 79–83. IEEE.

Tang, M., Q. Zhao, S. X. Ding, H. Wu, L. Li, W. Long, and B. Huang. 2020. “An Improved LightGBM Algorithm for Online Fault Detection of Wind Turbine Gearboxes.” *Energies* 13 (4): 807. https://doi.org/10.3390/en13040807.

Tang, R., L. De Donato, N. Besinović, F. Flammini, R. M. Goverde, Z. Lin, et al. 2022. “A Literature Review of Artificial Intelligence Applications in Railway Systems.” *Transportation Research Part C: Emerging Technologies* 140: 103679. https://doi.org/10.1016/j.trc.2022.103679.

Vyas, S., M. Shabaz, P. Pandit, L. R. Parvathy, and I. Ofori. 2022. “Integration of Artificial Intelligence and Blockchain Technology in Healthcare and Agriculture.” *Journal of Food Quality* 2022: 1–11. http://dx.doi.org/10.1155/2022/4228448.

Wang, Y., Z. Pan, J. Zheng, L. Qian, and M. Li. 2019. “A Hybrid Ensemble Method for Pulsar Candidate Classification.” *Astrophysics and Space Science* 364 (8): 1–13. https://link.springer.com/article/10.1007/s10509-019-3602-4.

Xu, H., Z. Yan, B. Ji, P. Huang, J. Cheng, and X. Wu. 2022. “Defect Detection in Welding Radiographic Images Based on Semantic Segmentation Methods.” *Measurement* 188: 110569. https://doi.org/10.1016/j.measurement.2021.110569.

Yan, Y., D. Liu, B. Gao, G. Y. Tian, and Z. C. Cai. 2020. “A Deep Learning-Based Ultrasonic Pattern Recognition Method for Inspecting Girth Weld Cracking of Gas Pipeline.” *IEEE Sensors Journal* 20 (14): 7997–8006. https://doi.org/10.1109/JSEN.2020.2982680.

Yuan, J., X. Zeng, H. Wu, W. Zhang, J. Zhou, and B. Chen. 2021. “Analytical Determination of Interwell Connectivity Based on Interwell Influence.” *Tsinghua Science and Technology* 26 (6): 813–20. http://dx.doi.org/10.26599/TST.2020.9010039.

Zhang, M., Y. Guo, Q. Xie, Y. Zhang, D. Wang, and J. Chen. 2022. “Defect Identification for Oil and Gas Pipeline Safety Based on Autonomous Deep Learning Network.” *Computer Communications* 195: 14–26. https://doi.org/10.1016/j.comcom.2022.08.001.

Zocca, V., G. Spacagna, D. Slater, and P. Roelants. 2017. *Python Deep Learning.* Birmingham: Packt Publishing.

APPENDIX

Nomenclature list

3D CNN Three-Dimensional Convolutional Neural Network
AdaBoost Adaptive Boosting
AI Artificial Intelligence
ANN Artificial Neural Network
ANN stands for Artificial Neural Network
BHP stands for Bottom Hole Pressure
BP Backpropagation
CNN Convolutional Neural Network
CRM stands for Capacitance-Resistance Model
DL Deep Learning
EFAST Extended Fourier Amplitude Sensitivity Test
EFB Exclusive Feature Bundling
GOR (Gas-Oil Ratio)
GOSS Gradient-based One-Side Sampling
GSA Global Sensitivity Analysis
IG Injector
LightGBM Light Gradient Boosting Machine
LightGBM Light Gradient Boosting Machine
LSTM Long Short-Term Memory
MAE Mean Absolute Erro
MCRM Multiple Component Relative Mobility
MSE Mean Squared Error
NTG net-to-gross
P Prodector
PED Production and Exploration Division
PSOC4IC Particle Swarm Optimization for Interwell Connectivity
RMSE Root mean square error