Gazi University
Journal of Science
PART A: ENGINEERING AND INNOVATION
http://dergipark.org.tr/gujsa

Optimizing Choke Size to Minimize Sand Production in Oil Wells: A Machine Learning Approach

Sunday AGBONS IGBINERE[1*] Ikponmwosa OHENHEN[1] Edobor Frankie CHRISTOPHER[1]

[1] *University of Benin, Benin City, Edo State, Nigeria*

Keywords

Sand Production
Machine Learning
SHAP
ANN
XGBoost
Random Forest
Optimization
Choke
Genetic Algorithm
Sand Cut

Abstract

This study investigates the relationship between operational parameters like choke size etc. and sand production in some oilfields, aiming to optimize efficiency while minimizing sand-related challenges. Data visualization revealed trends in sand cut behaviour under varying gross rate, net rate, and BS&W conditions. Three machine learning models artificial neural network (ANN), Random Forest (RF) and extreme gradient boosting (XGBoost) were developed, with XGBoost achieving the highest accuracy. Extreme gradient boosting (XGBoost) outperformed the others by achieving the highest R-squared value of 0.952 and the lowest mean absolute error (MAE) and mean squared error (MSE), demonstrating its superior accuracy in predicting sand cut values. Shapley additive exPlanations (SHAP) analysis highlighted key parameters like manifold pressure, gas rate, and BS&W in predicting sand cut. Optimization using the Mealpy Genetic Algorithm yielded an optimal configuration gross rate of 750blpd, sand cut of 0.25pptb, and BS&W of 5.5%. Sensitivity analysis emphasized monitoring separator pressure and gas rate. The findings demonstrate the potential of integrating machine learning and optimization to enhance decision-making, reduce risks, and improve production efficiency. Recommendations for implementation and future research are provided to ensure sustainable operations.

Cite

Igbinere, S.A., Ohenhen, İ., & Christopher, E.F. (2025). Optimizing Choke Size to Minimize Sand Production in Oil Wells: A Machine Learning Approach. *GU J Sci, Part A*, *12*(2), 541-561. doi:10.54287/gujsa.1669814

Author ID (ORCID Number)		**Article Process**	
0000-0003-1908-389X	Sunday AGBONS IGBINERE	**Submission Date**	03.04.2025
0000-0002-5938-8973	İkponmwosa OHENHEN	**Revision Date**	08.04.2025
0009-0000-5319-891X	Edobor Frankie CHRISTOPHER	**Accepted Date**	02.05.2025
		Published Date	30.06.2025

1. INTRODUCTION

Sand production is a pervasive challenge in oil and gas wells, particularly in poorly consolidated sandstone formations. It occurs when the mechanical strength of the reservoir rock is insufficient to withstand the stresses induced by fluid flow during production. Sand production can lead to severe operational and economic consequences, including erosion of surface and subsurface equipment, formation subsidence, sand accumulation in surface facilities, and increased maintenance costs due to frequent well interventions (Dickson et al., 2003; Completion Tech., 1995). These issues are exacerbated in high-rate production wells, where the fluid flow velocity exceeds the critical threshold required to mobilize sand particles. The causes of sand production are multifaceted, influenced by factors such as the degree of rock consolidation, pore pressure reduction, production rate, reservoir fluid viscosity, and water cut. For instance, younger, poorly consolidated formations, such as those found in the Gulf of Mexico, Nigeria, and the North Sea, are

*Corresponding Author, e-mail: agbons.igbinere@uniben.ed

particularly prone to sanding due to their low compressive strength (Completion Tech., 1997). Additionally, increased water production can destabilize the sand arch around perforations, further exacerbating sand production (Carlson et al., 1992). During production, reduction in pore pressure due to fluid withdrawal leads to increase of grain to grain effective stress and is directly proportional to the decrease in pore pressure, shear failure of the rock occurs when effective stress reaches the threshold value (Tiab & Donaldson, 2004).

To mitigate these challenges, various sand control methods have been developed, ranging from mechanical exclusion techniques (e.g., screens and gravel packs) to chemical consolidation and operational strategies like rate control and selective perforation. Among these methods is rate control, and it involves regulating the flow rate through choke size variations in order to help minimize sand production and has gained attention due to its simplicity and cost-effectiveness. By optimizing the choke size, operators can control the pressure drawdown and fluid velocity, thereby reducing the likelihood of sand creation and mobilization. However, traditional rate control methods often rely on empirical guidelines and trial-and-error approaches, which may not account for the complex interplay of reservoir properties, fluid dynamics, and operational constraints. This limitation underscores the need for a more systematic and data-driven approach to optimize choke size and minimize sand production.

When a well is first put into production, there is typically an initial surge in sand production due to the mobilization of perforation debris alongside the reservoir fluid. After reaching certain flow rates, a steady state is achieved, leading to transient sand production (Shabdirova et al., 2024). Extensive research has been conducted to model the onset of sand production, resulting in various analytical, numerical, and empirical models that predict the critical flow conditions that trigger sand production to varying extents (Morita et al., 1989; Kessler et al., 1993; Weingarten & Perkins, 2007; Wang & Dusseault, 2010; Han et al., 2011; Al-Shaaibi et al., 2013; Fuh & Morita, 2013; Wu et al., 2016; Papamichos & Furui, 2019). Laboratory experiments are instrumental in studying the mechanisms behind sand production, though their outcomes are highly influenced by boundary conditions, and may not always be directly applicable to field-scale scenarios (Skjaerstein et al., 1997; Nouri et al., 2006a; van den Hoek et al., 2007; Fattahpour et al., 2012; Wu et al., 2016; Kozhagulova et al., 2021; Shabdirova et al., 2022). These experimental results are primarily used to calibrate and validate analytical and numerical sand production models.

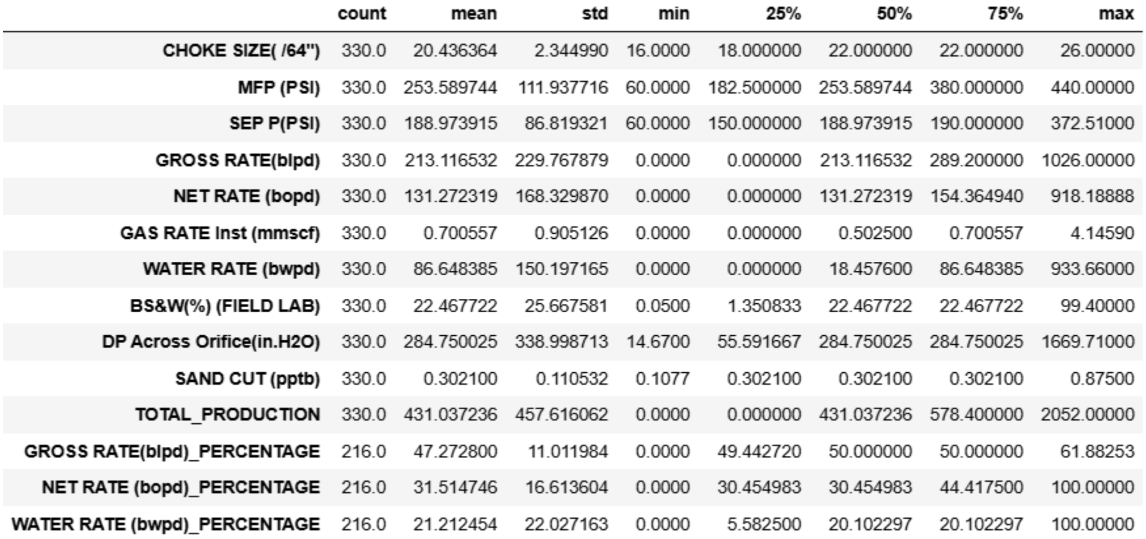
Despite significant progress in analytical, numerical, and experimental methods for predicting sand production, these approaches are frequently constrained by the need for simplifying assumptions, high computational demands, and difficulties in translating laboratory-scale results to real-world field conditions. In recent years, machine learning (ML) has emerged as a powerful tool for addressing complex problems in the oil and gas industry (Anirbib et al., 2021). ML algorithms, such as artificial neural networks (ANNs), support vector machines (SVMs) as well as fuzzy logic (FL) as classification and regression problem tools (Ani et al., 2016), and particle swarm optimization (PSO), have been successfully applied to predict sand production and optimize well performance. The prediction of sanding onset can be framed as a classification problem, with two possible outcomes: whether sanding will occur or not under specific conditions. Multiple studies have effectively utilized machine learning algorithms to predict sanding onset and evaluate the importance of various input variables. For example, Kanj and Abousleiman, (1999) pioneered the use of ANNs to predict sand onset, while subsequent studies have demonstrated the effectiveness of ML techniques in estimating critical drawdown pressure (CDDP) and identifying sand-prone zones (Azad et al., 2011; Khamehchi et al., 2014). Khamehchi et al. (2014) analyzed sand production onset in 23 field datasets from the North Adriatic Sea, using simple regression and artificial neural network (ANN) algorithms to predict critical total drawdown (CTD). Gharagheizi et al. (2017) applied support vector machine (SVM) to predict sanding onset using data from 31 wells in the Northern Adriatic Basin, with the algorithm demonstrating high predictive performance as assessed by the receiver operating characteristics (ROC) curve. Ketmalee and Bandyopadhyay (2018) utilized ANN to generate synthetic logs for sand prediction tools in the Bongkot field, Thailand, successfully validating the model with real well data. Ngwashi et al. (2021) compared the performance of ANN with backpropagation and SVM in predicting sanding onset in the Niger Delta region, achieving accuracy rates of 80% and 100%, respectively. Abdelghany et al. (2022) employed a probabilistic neural network to predict sand production onset by inferring reservoir properties, achieving higher accuracy compared to conventional models. Song et al. (2022) applied four machine learning algorithms to predict sand production in gas-hydrate sediments, recommending XGBoost for early-stage predictions. These advancements highlight the potential of ML to enhance sand management strategies by providing accurate, real-time predictions and optimizing operational parameters. This study proposes a machine learning approach to optimize choke size for minimizing sand production in oil wells. By leveraging historical production data, reservoir properties, and operational parameters, we aim to develop a predictive model that can identify the optimal choke size for a given well configuration. The proposed approach not only addresses the limitations of traditional rate control methods but also provides a scalable solution for improving well productivity and reducing sand-related costs. The integration of ML into sand management represents a significant step forward in the quest for sustainable and efficient oil and gas production as well as assisting in solving other reservoir engineering issues (Anifowose et al., 2017a, b).

2. MATERIAL AND METHOD

2.1. Data collection

The dataset used in this study was obtained from a producing asset located in the Niger Delta region. It comprises production data from eight wells spanning six different oil fields, each with varying geological and operational characteristics. The data was sourced from field operational logs and production monitoring systems, capturing a wide range of field conditions that reflect both stable and transient production phases. Key parameters recorded include: Gross Rate in barrel of liquid per day (blpd), Net Rate in barrel of oil per day (bopd), Sand Cut in parts per thousand barrels (pptb), BS&W in percentage (%), Gas Rate in million standard cubic feet (mmscf), manifold flowing pressure (MFP) in pounds per square inch (psi). The summary of the utilised dataset is shown in Table 1.

Table 1. *Summary Statistic of Data*



	count	mean	std	min	25%	50%	75%	max
CHOKE SIZE(/64")	330.0	20.436364	2.344990	16.0000	18.000000	22.000000	22.000000	26.00000
MFP (PSI)	330.0	253.589744	111.937716	60.0000	182.500000	253.589744	380.000000	440.00000
SEP P(PSI)	330.0	188.973915	86.819321	60.0000	150.000000	188.973915	190.000000	372.51000
GROSS RATE(blpd)	330.0	213.116532	229.767879	0.0000	0.000000	213.116532	289.200000	1026.00000
NET RATE (bopd)	330.0	131.272319	168.329870	0.0000	0.000000	131.272319	154.364940	918.18888
GAS RATE Inst (mmscf)	330.0	0.700557	0.905126	0.0000	0.000000	0.502500	0.700557	4.14590
WATER RATE (bwpd)	330.0	86.648385	150.197165	0.0000	0.000000	18.457600	86.648385	933.66000
BS&W(%) (FIELD LAB)	330.0	22.467722	25.667581	0.0500	1.350833	22.467722	22.467722	99.40000
DP Across Orifice(in.H2O)	330.0	284.750025	338.998713	14.6700	55.591667	284.750025	284.750025	1669.71000
SAND CUT (pptb)	330.0	0.302100	0.110532	0.1077	0.302100	0.302100	0.302100	0.87500
TOTAL_PRODUCTION	330.0	431.037236	457.616062	0.0000	0.000000	431.037236	578.400000	2052.00000
GROSS RATE(blpd)_PERCENTAGE	216.0	47.272800	11.011984	0.0000	49.442720	50.000000	50.000000	61.88253
NET RATE (bopd)_PERCENTAGE	216.0	31.514746	16.613604	0.0000	30.454983	30.454983	44.417500	100.00000
WATER RATE (bwpd)_PERCENTAGE	216.0	21.212454	22.027163	0.0000	5.582500	20.102297	20.102297	100.00000

2.1.1. Data quality and pre-processing

Before analysis, the dataset underwent rigorous quality checks to address missing values, outliers, and inconsistencies in measurement units. These pre-processing steps were essential to ensure data reliability for machine learning model training and optimization. To further understand the underlying data behaviour, the distribution of key variables—Sand Cut, Gross Liquid Rate, Water Rate, and BS&W—were visualized using histograms and kernel density estimation (KDE) plots (Figure 1-4). These visualizations revealed key distribution patterns and potential modelling implications. For instance, Gross Rate exhibited a right-skewed (log-normal) distribution, justifying the use of tree-based models like Random Forest or XGBoost, which are robust to non-Gaussian input features.

In particular, the distribution of Sand Cut (Figure 2) shows a sharp peak around 0.3 pptb, with a long tail extending toward higher values. This indicates that most wells operate within a narrow, low-sand range, while a few instances of higher sand production may reflect critical operational anomalies or edge cases. Marking a critical threshold of 0.5 pptb helps identify data points that exceed acceptable sand production limits and might require intervention or optimization of flow parameters. Upon data pre-processing, a Pearson correlation matrix was generated to examine the linear relationships among all numerical features in the dataset, as shown in Figure 5. This analysis aids in identifying potential multicollinearity and understanding how input variables interact. Notably, choke size (in 64ths of an inch) shows a negative correlation with sand cut (−0.23), suggesting that increasing the choke size may reduce sand production. This is consistent with fluid dynamics principles—larger choke openings reduce flow velocity, which can lower the mobilization of sand particles, aligning with Darcy's law. Additionally, net oil rate exhibits strong positive correlations with gross liquid rate (0.75) and gas rate (0.67), which reinforces its dependency on both fluid phases. Conversely, water rate and BS&W (Basic Sediment and Water) show a significant inverse relationship with oil production, which is critical for sand-prone reservoirs where higher water production can increase sand-related issues. Figure 2 showed a right-skewed pattern with most values clustered around 0.3 pptb. A critical threshold of 0.5 pptb is highlighted, beyond which sand production may pose operational risks.

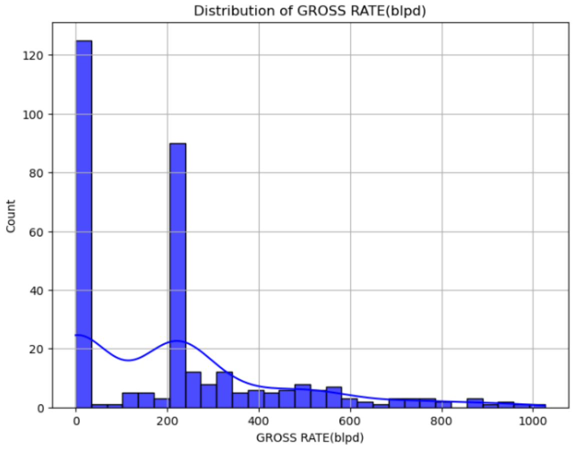


Figure 1. Distribution of Gross Rate

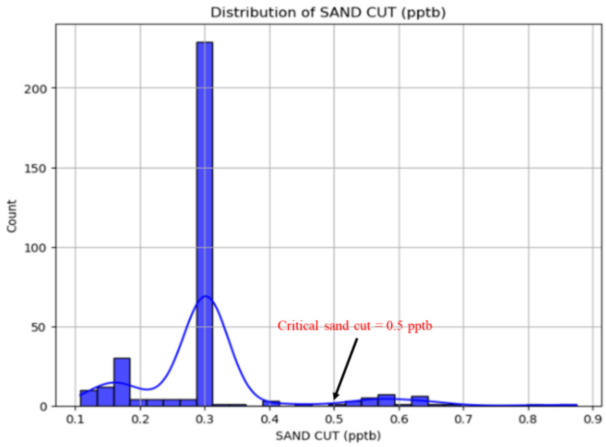


Figure 2. Distribution of Sand Cut

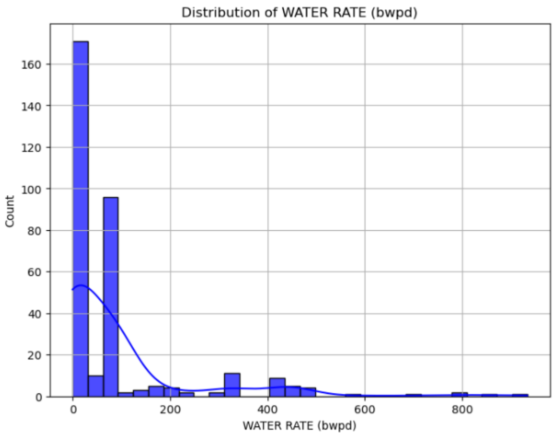


Figure 3. Distribution of Water Rate

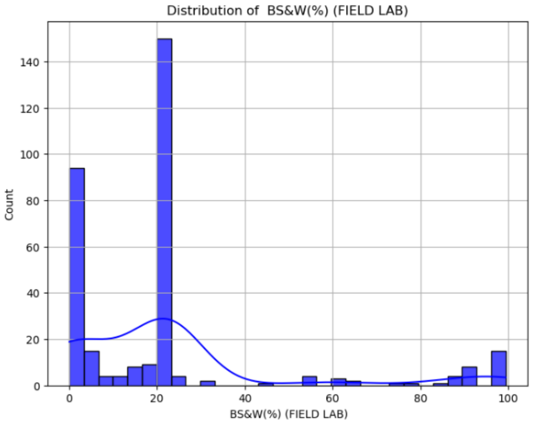
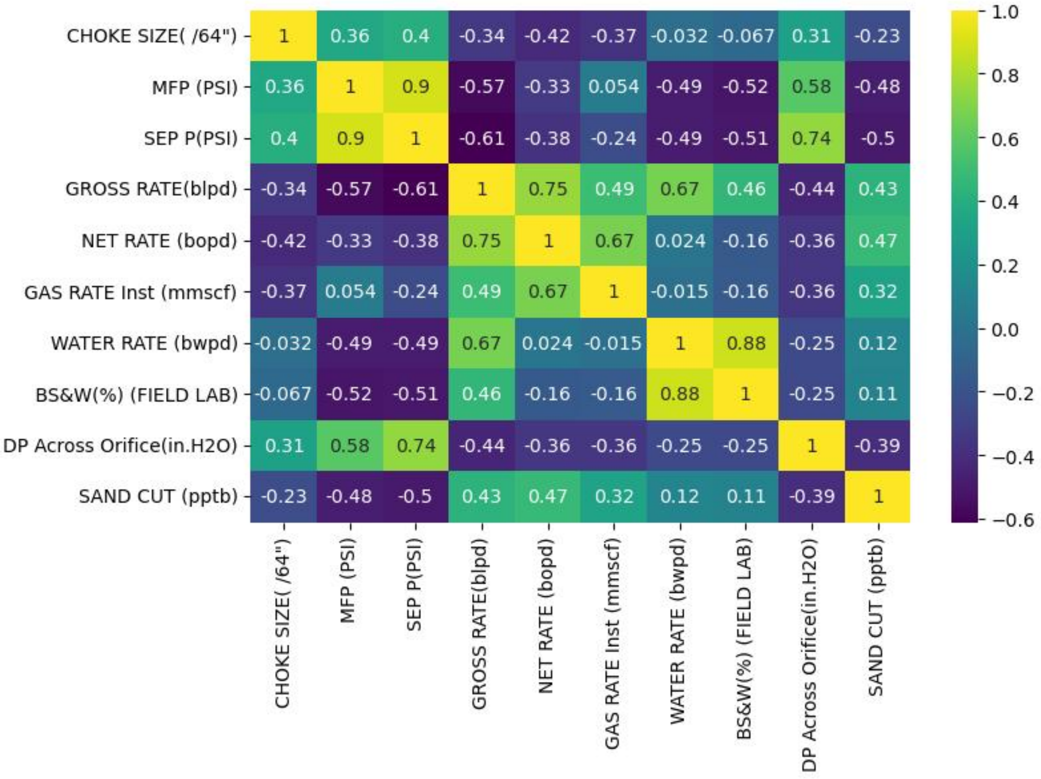


Figure 4. Distribution of BS&W



	CHOKE SIZE(/64")	MFP (PSI)	SEP P(PSI)	GROSS RATE(blpd)	NET RATE (bopd)	GAS RATE Inst (mmscf)	WATER RATE (bwpd)	BS&W(%) (FIELD LAB)	DP Across Orifice(in.H2O)	SAND CUT (pptb)
CHOKE SIZE(/64")	1	0.36	0.4	-0.34	-0.42	-0.37	-0.032	-0.067	0.31	-0.23
MFP (PSI)	0.36	1	0.9	-0.57	-0.33	0.054	-0.49	-0.52	0.58	-0.48
SEP P(PSI)	0.4	0.9	1	-0.61	-0.38	-0.24	-0.49	-0.51	0.74	-0.5
GROSS RATE(blpd)	-0.34	-0.57	-0.61	1	0.75	0.49	0.67	0.46	-0.44	0.43
NET RATE (bopd)	-0.42	-0.33	-0.38	0.75	1	0.67	0.024	-0.16	-0.36	0.47
GAS RATE Inst (mmscf)	-0.37	0.054	-0.24	0.49	0.67	1	-0.015	-0.16	-0.36	0.32
WATER RATE (bwpd)	-0.032	-0.49	-0.49	0.67	0.024	-0.015	1	0.88	-0.25	0.12
BS&W(%) (FIELD LAB)	-0.067	-0.52	-0.51	0.46	-0.16	-0.16	0.88	1	-0.25	0.11
DP Across Orifice(in.H2O)	0.31	0.58	0.74	-0.44	-0.36	-0.36	-0.25	-0.25	1	-0.39
SAND CUT (pptb)	-0.23	-0.48	-0.5	0.43	0.47	0.32	0.12	0.11	-0.39	1

Figure 5. Pearson Correlation Chart

2.2. Machine learning model development

Figure 6 presents the workflow adopted for machine learning model development in this study. The dataset was first imported into a Python environment (Jupyter Notebook), where it underwent several pre-processing steps. Data cleaning was performed to address missing values and eliminate inconsistencies, ensuring the quality of inputs fed into the models. The cleaned dataset was then partitioned into training and testing sets using a 70:30 split, allowing the models to learn patterns from the majority of the data while preserving a portion for unbiased evaluation. Feature scaling was applied using the Standard Scaler, particularly for models sensitive to the magnitude of input features such as Support Vector Machines (SVM) and Artificial Neural Networks (ANN). This step was omitted for tree-based models like Random Forest (RF) and Extreme Gradient Boosting (XGB), which are inherently robust to feature scaling.

Model development was carried out using SVM, ANN, RF, and XGB algorithms. To optimize each model's performance, hyperparameter tuning was conducted using GridSearchCV with five-fold cross-validation. This approach systematically searched through specified parameter combinations to identify configurations that yielded the best learning outcomes. After training, the models were integrated into a framework that enabled comparison and further interpretability through feature importance analysis using SHAP values, enhancing transparency in understanding which input variables most influenced the predicted sand cut.

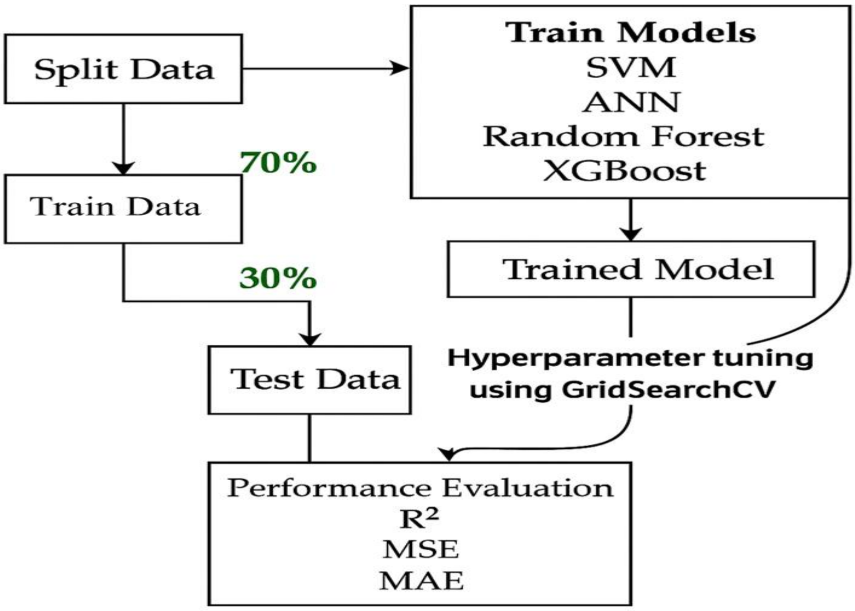


***Figure 6**. Machine Learning Workflow*

2.2.1. Random forest (RF)

Random Forest (RF) is a widely used machine learning technique for building predictive models across various research domains. It operates on the principle of bagging, which combines bootstrapping and aggregation to enhance model performance (Ali et al., 2020). A key objective in predictive modelling is often to minimize the number of variables required for accurate predictions, thereby reducing data collection efforts and improving efficiency (Speiser et al., 2019). In RF, each decision tree is trained on a randomly selected subset of input data (a bootstrap sample) and develops independently. The model's accuracy stems from the collective predictions of these individual trees, leveraging their combined strength. The RF process involves three main steps: (1) generating *n* bootstrap samples from the input variables, (2) constructing unpruned regression trees using the optimal predictor split, and (3) aggregating predictions from the *n* trees (Nosakhare et al., 2024).

For a random data subset *d(x, y)* the decision tree iteratively partitions the variable space *x* as samples with close targets are aggregated. The data at node m is represented by $d_m$with N_m samples, and d_m may be partitioned into subsets denoted as d_m^{right} and d_m^{left}using each candidate split $\theta(p, t_m)$ where p and t_m refers to a feature and threshold, respectively as shown in equation 1.

$$\begin{cases} d_m^{left} = \{(x,y) |\, x_p \leq t_m\} \\ d_m^{right} = d_m / d_m^{left} \end{cases} \tag{1}$$

The candidate split is defined as a loss function $H(\bullet)$, as shown in equation 2.

$$H(d_m) = \frac{1}{N_m} \sum_{y \in d_m} (y - \bar{y}_m)^2 \tag{2}$$

Where $\bar{y}_m = \frac{1}{N_m} \sum_{y \in d_m}$

The minimization of Equation (2) yields the parameters $\theta(p, t_m)$.

$$G(d_m, \theta) = \frac{N_m^{left}}{N_m} H\left(d_m^{left}(\theta)\right) + \frac{N_m^{right}}{N_m} H\left(d_m^{right}(\theta)\right) \tag{3}$$

The recursion of equation 3 continues for d_m^{right} and d_m^{left} until the maximum depth is achieved. The prediction of the RF model is thus obtained from equation 4.

$$f(x) = \frac{1}{K} \sum_{K=1}^{K} DT_i(x) \tag{4}$$

where K is the number of decision trees (DT) in the random forest.

2.2.2 Extreme gradient boosting (XGB)

Extreme Gradient Boosting (XGBoost) is a highly scalable and efficient tree-based machine learning algorithm widely adopted across diverse data analysis fields. As an advanced implementation of gradient boosting, XGBoost excels in both regression and classification tasks. Extreme gradient boosting is a type of boosting tree that has been used to determine nonlinear and local relationship in sand production prediction (Song et al., 2022). Its foundation lies in the boosting principle, which integrates the predictions of multiple weak learners through an additive training process to create a strong, accurate model. This approach not only boosts predictive performance but also reduces overfitting caused by insufficient data, small learning rate and large tree depth that captured noise rather than generalizable trends. XGB also enhances computational efficiency (Alshboul et al., 2022; Sheridan et al., 2016). The general function of the forecasting is set up at step p, as presented in equation 5

$$f_i^{(p)} = \sum_{k=1}^{p} f_k\,(x_i) = f_i^{(p-1)} + f_p\,(x_i) \tag{5}$$

where $f_p\,(x_i)$ denotes the learner at step p, $f_i^{(p)}$denotes the prediction at p, $f_i^{(p-1)}$denotes the prediction at $p-1$, and x_i denotes the input features.

To balance overfitting while maintaining computational efficiency, XGBoost employs a refined analytical formula, as shown in equation 6, to assess the model's "goodness of fit" to the original function. This method optimizes performance by controlling model complexity and improving predictive accuracy.

$$Objective^{(p)} = \sum_{k=1}^{n} l\,(\bar{y}_i\,, y_i) + \sum_{k=1}^{p} \sigma(f_i) \tag{6}$$

where l presents the loss function, n presents the number of observations utilized, and σ presents the regularization term as represented in equation 7.

$$(f) = \boldsymbol{\Theta} T + 0.5\lambda\omega^2 \tag{7}$$

where ω expresses vector scores in leaves, Υ expresses the minimal loss necessary to divide the leaf node further, and λ expresses the regularization parameters.

2.2.3. Artificial neural network (ANN)

ANN is a computational model inspired by the structure and function of the biological neural networks that make up the brain. The ANN model used for this study was a feed-forward neural network consisting of n hidden layers each with a set of neurons that can estimate a given output response (y) from provided input data (x). Figure 7 illustrates the architecture of the Artificial Neural Network (ANN) model used in this study. The network consists of an input layer, four hidden layers, and a single output layer. Each hidden layer utilizes the ReLU (Rectified Linear Unit) activation function to introduce non-linearity and enhance model learning. The number of hidden layers and their respective neuron configurations were selected through hyperparameter optimization using GridSearchCV, aiming to achieve an optimal balance between model complexity and generalization performance. Dropout layers were not incorporated, as the dataset was sufficiently regularized through tuning and early stopping.

The mathematical concept of ANN involves a set of interconnected nodes or neurons that process and transmit information. Each neuron calculates a linear combination (z) of n input features (x_i) with corresponding weight (w_i) and bias (b) as shown in equation 8. The weights of the connections between the neurons are adjusted during training to minimize the error between the predicted output and the true output. The selection of the right activation function is very essential in the ANN modelling process as it is chiefly responsible for the mapping of the inputs to the outputs.

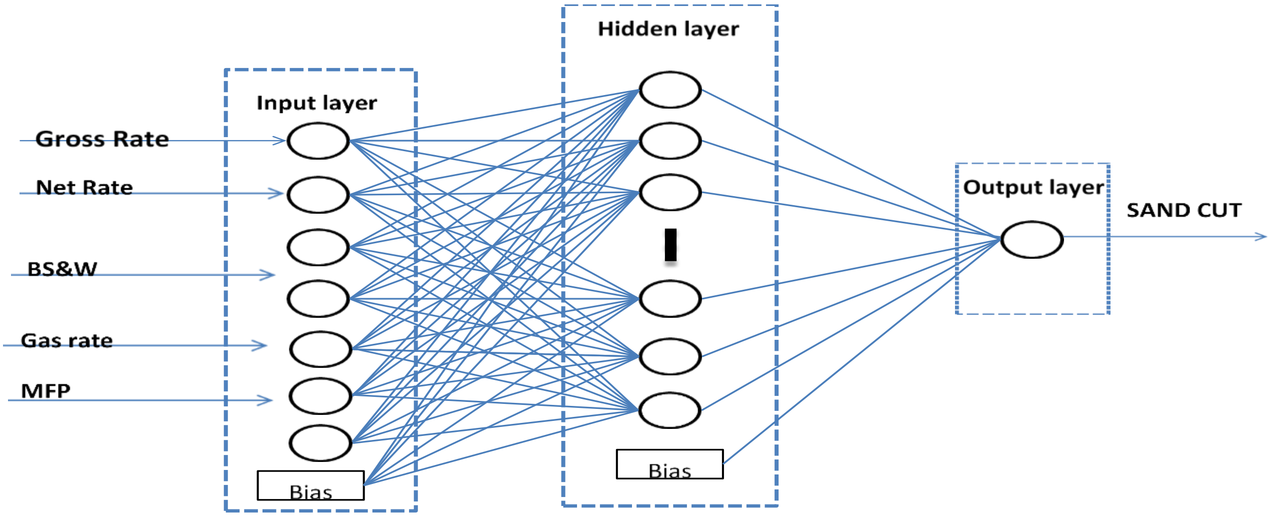


Figure 7. *ANN Architecture for the Implemented Dataset*

$$z = \sum_{i=1}^{n} x_i w_i + b \tag{8}$$

During the ANN process, the transformed version of the output data is forwarded from one hidden layer to another. The network is trained through an updating process that tunes the model weights to minimize model loss (L) which is typically taken as the root mean square error (RMSE) as shown in equation 9, where $\hat{y}_i$ is the model prediction. The minimization is done through a gradient descent method where the weights are manipulated as a function of L and scaled using the learning rate (λ) as shown in equation 10.

$$L = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{9}$$

$$w_{t+1} = w_t - \lambda \frac{\partial \mathrm{L}}{\partial w_t} \tag{10}$$

To avoid overfitting, a k-fold cross-validation was carried out using the output dataset. The ANN model was implemented in Tensor Flow version 2.11.0 library. The ANN hyper parameters were optimized using a specially created Python script that compared the prediction accuracy of different networks trained with different combinations of hyper parameters.

2.2.4. Genetic algorithms

Adaptive heuristic search algorithms, or genetic algorithms (GAs) (Figure 8), make up the majority of evolutionary algorithms. Natural selection and genetics serve as the foundation for genetic algorithms. These are clever uses of sporadic searches that are aided by past data to focus the search on areas of the solution space where performance is higher. They are frequently employed to produce excellent solutions for search and optimization issues.

Because genetic algorithms mimic the process of natural selection, species that are able to adapt to changes in their surroundings have a better chance of surviving, procreating, and evolving into new generations. To put it another way, they mimic the "survival of the fittest" among people from successive generations in order to address an issue. Every generation is made up of a population of individuals, and each individual stands for a potential solution or point in the search space. Every person is represented as a string made up of bits, characters, integers, and floats. The chromosome is comparable to this string.

2.3. SHapley additive explanations (SHAP)

SHAP (SHapley Additive exPlanations) is a unified framework for interpreting machine learning model outputs. It is based on game theory and assigns each feature an importance value for a particular prediction, offering insights into the contribution of individual input features to the model's predictions. This study employs SHAP to enhance the explainability of the machine learning models used for predicting sand cut.

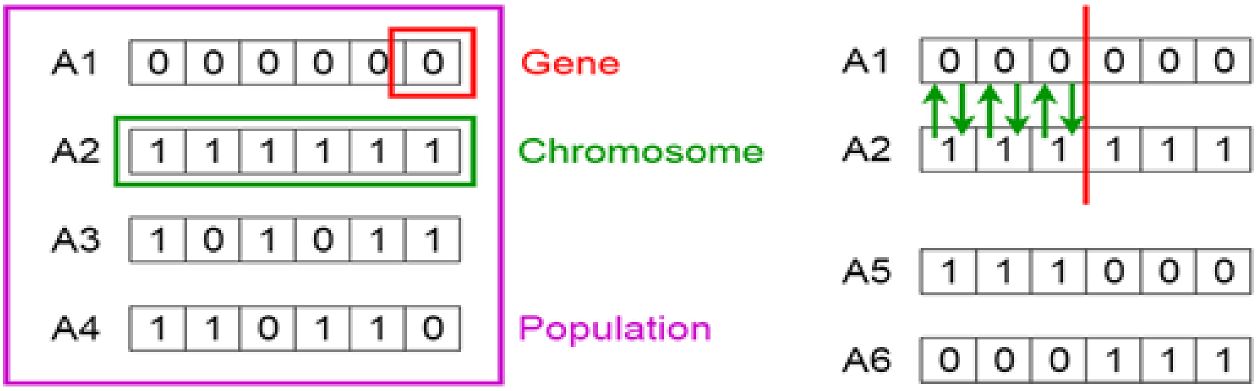


Figure 8. *Simple Implementation of Genetic Algorithms*

2.3.1. Theoretical framework

SHAP leverages concepts from cooperative game theory, particularly Shapley values, to fairly distribute the “payout” (i.e., prediction) among the “players” (i.e., input features). Each feature’s contribution is calculated by considering all possible subsets of features, ensuring a comprehensive assessment of its impact on the prediction. The main properties of SHAP values include:

1. **Local Accuracy:** The sum of SHAP values for all features equals the model’s prediction.

2. **Consistency:** If a model changes such that a feature contributes more to predictions, its SHAP value will not decrease.

3. **Missingness:** Features that do not influence the prediction will have a SHAP value of zero.

2.3.2. Implementation in this study

In this research, SHAP was applied to interpret the outputs of the three machine learning models namely; artificial neural Network (ANN), random forest (RF), and extreme gradient boosting (XGBoost). The SHAP framework was particularly valuable for:

1. **Identifying Key Features:** Determining the most influential parameters affecting sand production.

2. **Feature Interaction:** Exploring how different features interact and contribute to predictions.

3. **Global and Local Interpretations**: Understanding the overall behaviour of the model (global) and the specific reasoning behind individual predictions (local).

3. RESULTS AND DISCUSSION

The field production data was used to train multiple machine learning (ML) models, optimizing their respective hyperparameters to enhance predictive performance. Table 1 presents the optimal hyperparameters obtained for each model through a combination of 5-fold cross-validation and a full-factor experimental grid search. The performance of these models was evaluated using key statistical metrics, which are discussed in section 3.2.

3.1. Hyperparameter tuning

To ensure optimal model performance, Grid Search CV with 5-fold cross-validation was employed for hyperparameter tuning. The best-performing hyperparameters for each model are summarized in Table 2.

3.2. Model performance evaluation

Presented in Table 3 are the mean absolute error (MAE), mean squared error (MSE) and the R-squared or coefficient of determination (R^2), and these parameters helps to determine the differences between actual and predicted values as well as how well regression model explains the variance in the dependent variables. Among the models shown in Table 3, extreme gradient boosting (XGBoost) demonstrated the best performance, with the lowest values for mean absolute error and mean squared error as well as the highest R^2, indicating its superior predictive capability for the dataset. Figure 9 and 10 showed the actual and predicted sand cut values using the extreme gradient boosting model and the artificial neural network model.

3.3. Sand production optimization

The relationship between net rate (bopd) and sand cut (pptb), as shown in Figure 11, reveals two distinct regimes. At lower net rates (<500 bopd), sand cut exhibits significant variability, likely due to unstable sand arches in the near-wellbore region that collapse intermittently under subcritical flow velocities (Tixier, 1949). However, beyond 500 bopd, sand cut stabilizes (Figure 11), suggesting that exceeding this critical flow velocity minimizes sand arch collapse and promotes consistent sand transport. This stabilization implies that maintaining net rates above 500 bopd (shaded as the ‘Safe Zone’ in Figure 11) could reduce unpredictability in sand production, offering a practical threshold for operational optimization. Larger choke openings increase production rates (and vice versa), enabling operators to target the stabilized ‘Safe Zone’ for sand management. While higher net rates improve predictability, further analysis is required to balance sand mitigation with reservoir pressure sustainability. These findings align with field observations where sustained high-rate production often correlates with reduced sand-related downtime.

3.4. SHAP sensitivity analysis

The SHAP analysis (Figure 12–13) quantifies the relative importance and directional effects of operational parameters on sand mitigation. Key findings include:

1. **Prioritization of Critical Parameters**:

- MFP (psi) emerged as the most influential parameter (SHAP value: +0.04), exerting 4× greater impact than Choke Size (/64") (SHAP value: +0.01) (Figure 12). This aligns with field observations where pressure fluctuations correlate strongly with sand ingress.
- Practical Insight: To mitigate sand production, operators should prioritize real-time MFP monitoring and stabilization protocols over secondary adjustments like choke size optimization.

2. **Directionality of Feature Impact**:

- The beeswarm plot (Figure 13) reveals that higher MFP values (associated with stable reservoir conditions) consistently reduce sand cut, whereas elevated gas rates correlate with increased sand risk.
- Field Action: Maintaining optimal MFP thresholds (e.g., within ±5% of baseline) is recommended to suppress sand activity, while gas rate surges should trigger mitigation workflows (e.g., choke throttling, screen inspections).

3. **Interaction Effects (Future Work)**:
While the current SHAP analysis focuses on individual parameter impacts, the reviewer rightly highlights the need to explore **interaction effects** (e.g., how MFP and gas rate jointly influence sand dynamics). Although partial dependence plots (PDPs) are beyond the scope of this study, future work will integrate PDPs to quantify these synergies and refine multi-variable control strategies.

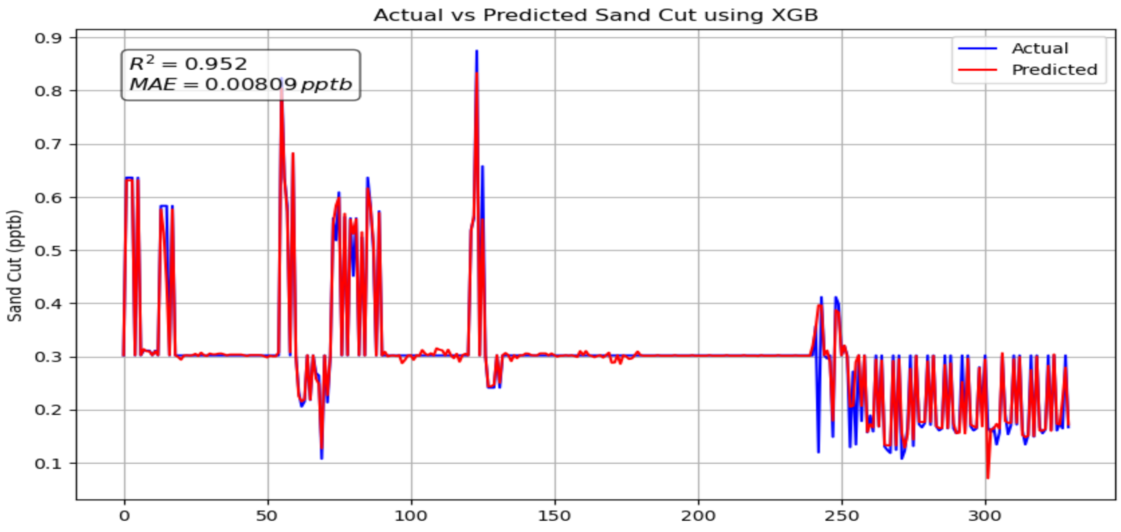


***Figure 9**. Actual vs Predicted Sand Cut Values using XGBoost*

3.5. Optimization using genetic algorithm (GA)

The Mealpy Genetic Algorithm (GA) was applied to optimize operational parameters to minimize sand production while maintaining production targets. The optimization objectives included maximizing gross rate and minimizing sand cut and BS&W.

***Table 2**. Optimal Hyperparameters for the Utilised Machine Learning Models*

Model	Hyperparameters	Final Optimized Value
ANN	Kernel_initializer	Normal
	model optimizer	Lbfgs
	Epochs	100
	layer 1 number of neurons	32
	layer 1 activation function	Relu
	layer 2 number of neurons	64
	layer 2 activation function	Relu
	layer 3 number of neurons	16
	layer 3 activation function	Relu
	layer 4 number of neurons	8
	layer 4 activation function	Relu
	output layer number of neurons	1
	output activation function	Relu
RF	n estimators	500
	max depth	5
	max features	3
XGB	max depth	5
	ϒ	0.01
	n estimators	500
	Subsample	1
	random state	9

***Table 3**. Model Performance Metrics*

Model	MAE	MSE	R^2
ANN	0.0154	0.0011	0.923
RF	0.0114	0.00086	0.942
XGBoost	0.0081	0.00058	0.952

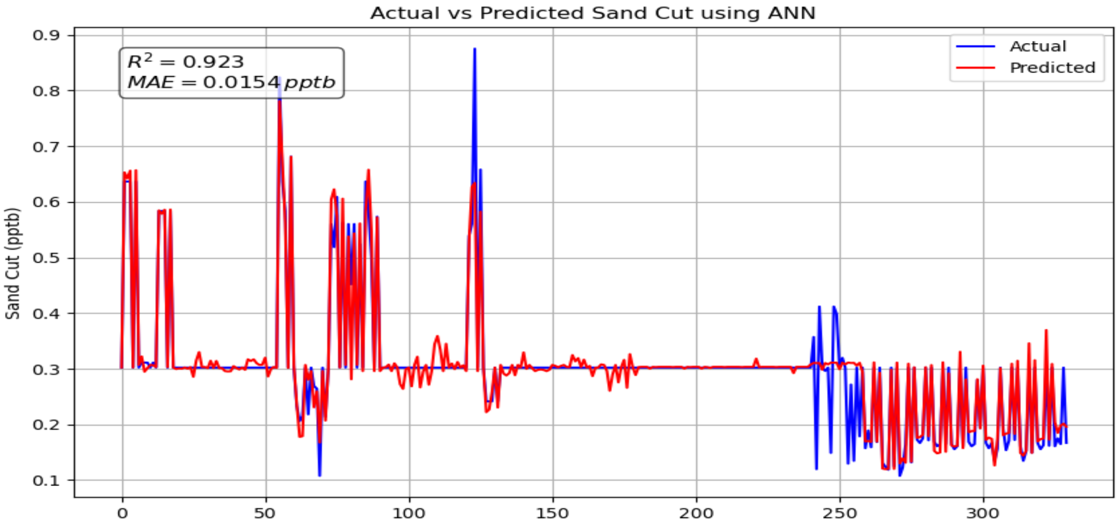


***Figure 10**. Actual vs Predicted Sand Cut Values using ANN*

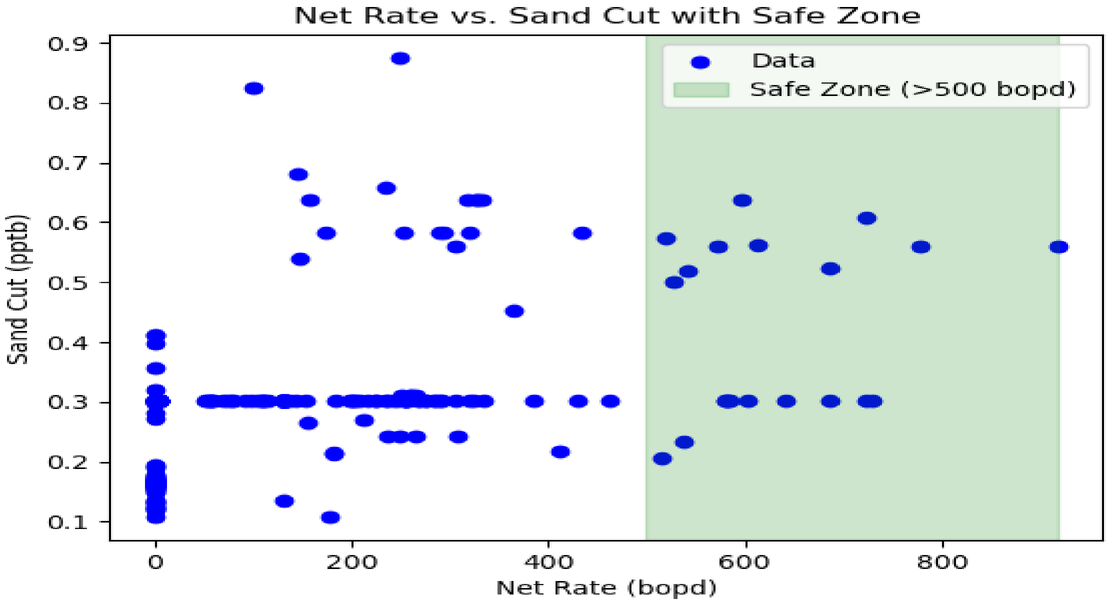


Figure 11. *Sand Production Prediction*

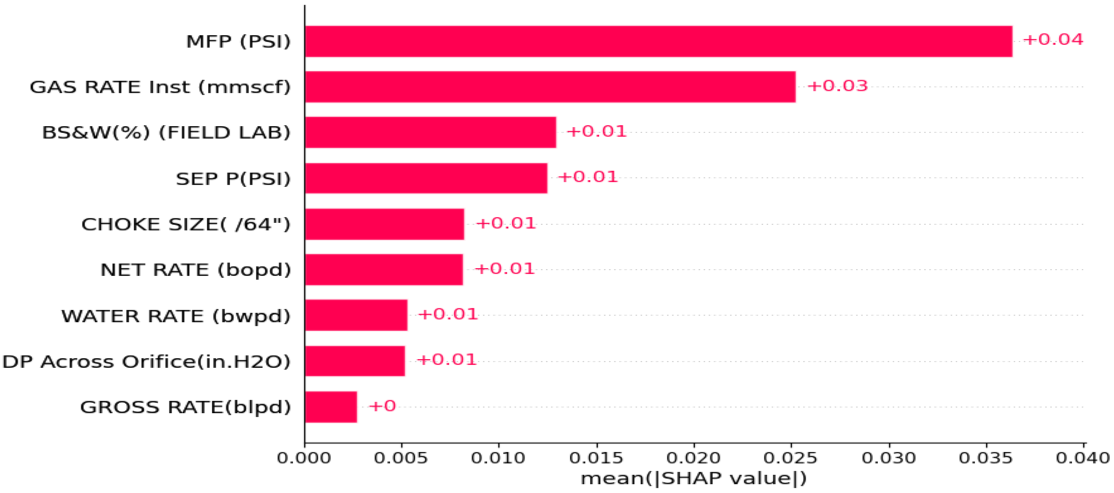


Figure 12. *Feature importance plot using SHAP*

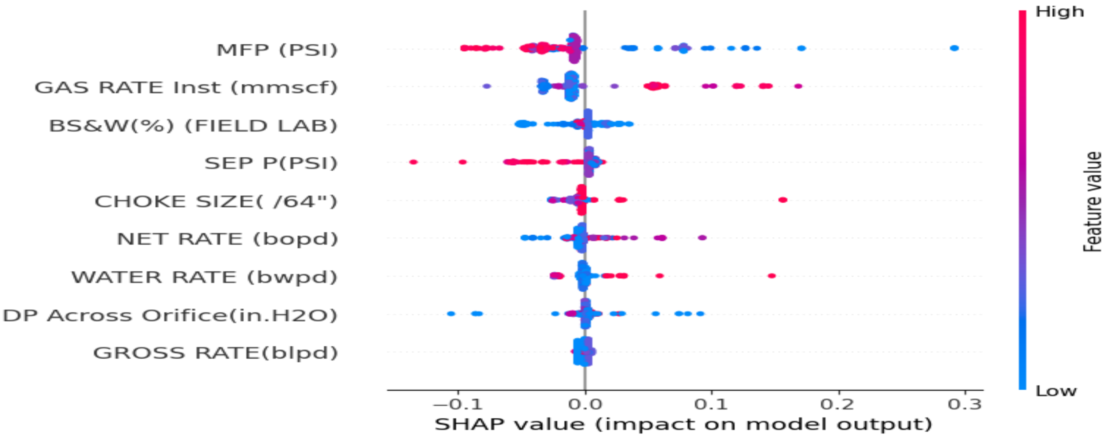


Figure 13. *Summary plot of model impact on sand cut using SHAP*

3.5.1. Optimization results

Table 4 shows the optimal parameters of the optimization algorithm used in minimising sand cut produced.

Table 4*. Optimal Parameters of the Genetic Algorithm*

Optimization Algorithm	Hyperparameters	Final Optimized Value
Genetic Algorithm	Epoch	100
	population size	100
	Pc	0.75
	Pm	0.25

The GA achieved an optimal configuration of parameters, yielding: optimal net rate of 750blpd minimized sand cut of 0.197pptb and minimized BS&W of 5.5%. The results indicate that GA effectively balances production efficiency with operational safety, reducing risks associated with excessive sand production. The genetic algorithm's convergence is shown in Figure 14, where the fitness value (representing sand mitigation efficacy) improves steadily over generations. This indicates robust optimization toward operational parameters that minimize sand production. The stabilization of fitness after ~40 generations suggests diminishing returns beyond this point, aligning with field observations that sand management gains plateau once critical thresholds (e.g., net rate >500 bopd) are sustained.

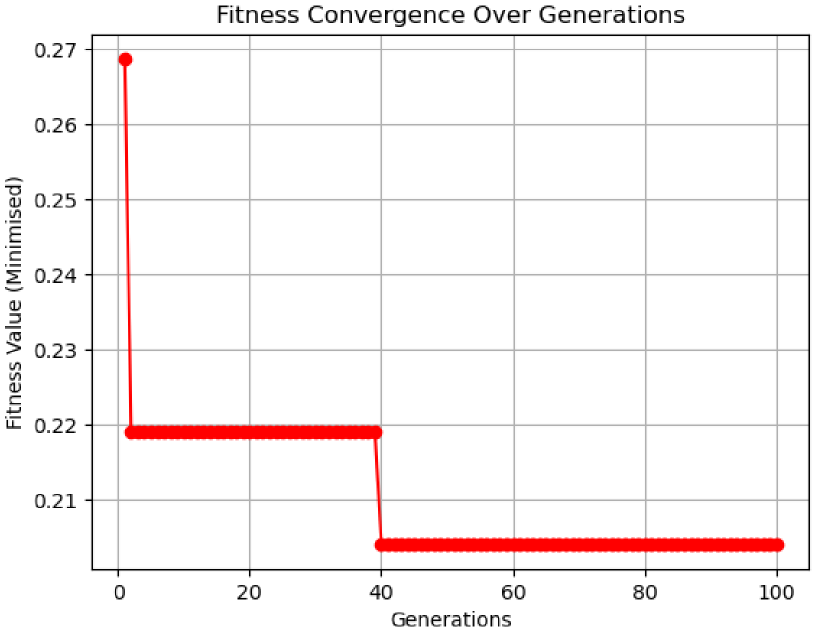


Figure 14. *Fitness convergence over 100 generations, demonstrating optimization progress toward maximizing sand mitigation*

3.6. Comparative analysis with previous studies

To further validate the performance of the developed XGBoost model, its predictive accuracy was compared with those reported in existing literature as shown in Figure 15. The XGBoost model in this study achieved an R^2 of 0.952, which is notably higher than those obtained in similar works. For example, Ketmalee and

Bandyopadhyay (2018) reported an R^2 of 0.908 using an Artificial Neural Network (ANN), while Ngwashi et al. (2021) also employed ANN but recorded a lower R^2 of 0.800. Song et al. (2022), who used a boosting tree algorithm similar to XGBoost, reported an R^2 of 0.940, slightly below the performance achieved in this study. Furthermore, Shabdirova et al. (2022) applied a Support Vector Machine (SVM) model and obtained an R^2 of 0.870. The superior performance of the present model can be attributed to the combination of advanced feature selection through SHAP analysis and hyperparameter tuning via GridSearchCV, as well as the integration of a Genetic Algorithm for optimization. These methodological enhancements have resulted in a more robust and reliable model capable of capturing the complex nonlinearities associated with sand production in oilfield operations.

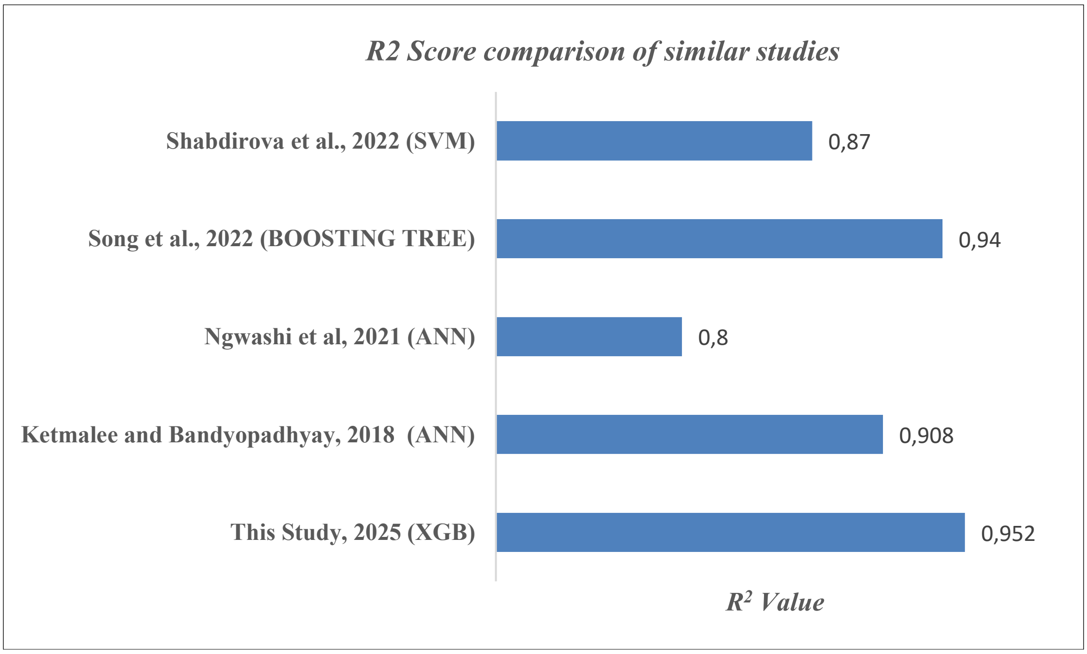


Figure 15. *Comparison of R^2 values obtained and selected related works in sand production prediction*

4. CONCLUSION

This study developed and validated an optimized machine learning framework to predict and minimize sand production in oilfield operations, with a focus on improving operational decision-making. Utilizing a multi-feature dataset obtained from eight wells across six different fields in the Niger Delta, the study implemented multiple supervised learning algorithms and employed advanced optimization techniques.

Among the models tested, Extreme Gradient Boosting (XGBoost) delivered the best performance, achieving an R^2 of 0.952, a mean absolute error (MAE) of 0.00809 pptb, and a mean squared error (MSE) of 0.00058 pptb^2. These results indicate its strong ability to capture complex nonlinear relationships inherent in sand production dynamics. Feature importance analysis conducted using SHAP values identified manifold flowing pressure (MFP), gas rate, and basic sediment and water (BS&W) as the most critical features influencing sand cut. These insights offer actionable guidance for production optimization and operational planning.

Furthermore, optimization using a Genetic Algorithm (GA) yielded a gross production rate of 750 blpd, BS&W of 5.5%, and minimized sand cut of 0.197 pptb as the ideal operational conditions. Sensitivity analysis further reinforced the critical role of MFP and gas rate, emphasizing the necessity of real-time monitoring to maintain effective sand control strategies. Consequently, the following inferences were arrived at from this study;

1. Three machine learning models, artificial neural network (ANN), random forest (RF) and extreme gradient boosting (XGBoost) were developed and evaluated.
2. Extreme gradient boosting (XGBoost) outperformed the others by achieving the highest R-squared value of 0.952 and the lowest mean absolute error (MAE) and mean squared error (MSE), demonstrating its superior accuracy in predicting sand cut values.
3. Feature importance analysis using SHAP values identified manifold flowing pressure (MFP), gas rate and BS&W are the most influential factors affecting sand production.
4. Optimization using a Genetic Algorithm (GA) successfully balanced production efficiency and sand control, yielding an optimal gross rate of 750blpd, minimized sand cut of 0.197pptb, and a reduced BS&W of 5.5%.
5. Sensitivity analysis further highlighted the critical role of manifold flowing pressure and gas rate, emphasizing the need for real-time monitoring to sustain effective sand control strategies optimal operational conditions.

Limitations and Future Work

- The dataset is limited to a specific set of wells within the Niger Delta region; expanding the model to include data from other regions and reservoirs would enhance generalizability.
- The integration of real-time sensor data is recommended to improve prediction accuracy and support dynamic decision-making.
- Future research should explore the use of hybrid and ensemble models, as well as incorporating domain-specific constraints and physics-based knowledge into machine learning architectures for more robust and interpretable predictions.

AUTHOR CONTRIBUTIONS

The three authors of this research study were involved in all the different stages of the work, and also contributed immensely in the reporting of this article for publication.

ACKNOWLEDGEMENT

The authors of this work acknowledge the immense contributions of Mr. Ifeanyi Cyriacus Azubuike and Mr. Shedrach Ikpeminoghena Igemhokhai of the Department of Petroleum Engineering, University of Benin, Benin City during the research stages of this work.

CONFLICT OF INTEREST

The authors of this research article copiously and clearly state that there is no conflict of interest before the conceptualization of this work, during the different stages involved in the study and after the completion of the work.

REFERENCES

Abdelghany, W. K., Hammed, M. S., Radwan, A. E., & Nassar, T. (2022). Implications of machine learning on geomechanical characterization and sand management: A case study from Hilal Field, Gulf of Suez, Egypt. *Journal of Petroleum Exploration and Production Technology*. https://doi.org/10.1007/s13202-022-01551-9

Ali, M., Prasad, R., Xiang, Y. & Yaseen, Z. M. (2020). Complete Ensemble Empirical Mode Decomposition Hybridized with Random Forest and Kernel Ridge Regression Model for Monthly Rainfall Forecasts, *J. Hydrol*. 584(2020) 124647, https://doi.org/10.1016/J.JHYDROL.2020.124647

Al-Shaaibi, S. K., Al-Ajmi, A. M., & Al-Wahaibi, Y. (2013). Three dimensional modeling for predicting sand production. Journal of Petroleum Science and Engineering, 109, 348–363.

Alshboul, O., Shehadeh, A., Almasabha, G., & Almuflih, A. S. (2022). Extreme Gradient Boosting-Based Machine Learning Approach for Green Building Cost Prediction. *Sustainability*, 14(11), 6651. https://doi.org/10.3390/su14116651

Ani, M., Oluyemi, G., Petrovski, A., & Rezaei-Gomari, S. (2016). Reservoir Uncertainty Analysis: the Trends from Probability to Algorithms and Machine Learning. *SPE Intelligent Energy International Conference and Exhibition*, 6-8 September 2016, Aberdeen, UK.

Anifowose, F. A., Labadin, J. & Abdulraheem, A. (2017a). Hybrid Intelligent Systems in Petroleum Reservoir Characterization and Modeling: The Journey so Far and the Challenges Ahead. *Journal of Petroleum Exploration and production Technology*, **7**(1) (2017), 251-263.

Anifowose, F. A., Labadin, J. & Abdulraheem, A. (2017b). Ensemble Machine Learning: An Untapped Modeling Paradigm for Petroleum Reservoir Characterization. *J. Petrol. Sci. Eng*., 151(2017), 480-487.

Anirbid, S., Kriti, Y., Kamakshi, R., Namrata, B. & Hemangi, O. (2021). Application of Machine Learning and Artificial Intelligence in Oil and Gas Industry. *Petroleum Research*, 6(4), 379-391. https://doi.org/10.1016/j.ptlrs.2021.05.009

Azad M., Zargar G. & Arabjamaloei, R. (2011). A New Approach to Sand Production Onset Prediction Using Artificial Neural Networks. *Petroleum Science and Technology*, 29:19, 1975-1983. https://doi.org/10.1080/10916460903551081

Carlson, J., Gurley, D., King, G., Price-Smith, C., & Waters, F. (1992). Sand Control: Why and How. *Oilfield Review;* Netherlands, *4*(4).

Completion Technology for Unconsolidated Formations, Revision 2, (1995) Completion Technology for Unconsolidated Formations, Revision 3, (1997). https://fr.scribed.com

Dickson, R. L., Raymond, C. A., Joe, W., & Wilkinson, C. A. (2003). Segregation of Eucalyptus Dunnii Logs using Acoustics. *Forest Ecology and Management*, *179*(1-3), 243-251.

Fattahpour, V., Moosavi, M., & Mehranpour, M. (2012). An experimental investigation on the effect of rock strength and perforation size on sand production. *Journal of Petroleum Science and Engineering, 86–87*, 172–189.

Fuh, G., & Morita, N. (2013). Sand production prediction analysis of heterogeneous reservoirs for sand control and optimal well completion design. *International Petroleum Technology Conference*. https://doi.org/10.2523/16940-MS

Gharagheizi, F., Mohammadi, A., Arabloo, M., & Shokrollahi, A. (2017). Prediction of sand production onset in petroleum reservoirs using a reliable classification approach. *Petroleum, 3*(2), 280–285. https://doi.org/10.1016/j.petlm.2016.02.001

Han, G., Shepstone, K., Harmawan, I., Er, U., Jusoh, H., Lin, L. S., Pringle, D., et al. (2011). A comprehensive study of sanding rate from a gas field, from reservoir to completion, production, and surface facilities. *SPE Journal, 16*(2), 463–481.

van den Hoek, P. J., Hertogh, G. M. M., Kooijman, A. P., de Bree, Ph., Kenter, C. J., & Papamichos, E. (2007). A new concept of sand production prediction, theory and laboratory experiments. *SPE Drilling & Completion, 15*(4), 261–273.

Kanj, M. Y. & Abousleiman, Y. (1999). Realistic Sanding Predictions: A Neural Approach. Paper Presented at the *SPE Annual Technical Conference and Exhibition*, October 3–6, 1999. SPE-56631-MS. https://doi.org/10.2118/56631-MS

Kessler, N., Wang, Y., & Santarelli, F. J. (1993). A simplified pseudo 3D model to evaluate sand production risk in deviated cased holes. *SPE Annual Technical Conference and Exhibition*.

Ketmalee, T., & Bandyopadhyay, P. (2018). Application of neural network in formation failure model to predict sand production. In *Offshore Technology Conference Asia 2018* (pp. 1–10). https://doi.org/10.4043/28506-ms

Khamehchi, E. Kivi, I. R. & Akbari, M. (2014). A Novel Approach to Sand Production Prediction using Artificial Intelligence. *Journal of Petroleum Science and Engineering*. 123(2014), 147-154. https://doi.org/10.1016/j.petrol.2014.07.033

Kozhagulova, A., Shabdirova, A., Minh, N. H., & Zhao, Y. (2021). An integrated laboratory experiment of realistic diagenesis, perforation and sand production using a large artificial sandstone specimen. *Journal of Rock Mechanics and Geotechnical Engineering*.

Morita, N., Whitfill, D. L., Fedde, O. P., & Lovik, T. H. (1989). Parametric study of sand-production prediction: Analytical approach. *SPE Production Engineering, 4*(1), 25–33. https://doi.org/10.2118/16990-PA

Ngwashi, A. R., Ogbe, D. O., & Udebhulu, D. O. (2021). Evaluation of machine-learning tools for predicting sand production. In *SPE Nigeria Annual International Conference and Exhibition 2021, NAIC 2021*, 1–16. https://doi.org/10.2118/207193-MS

Nosakhare, A., Igemhokhai, S., Aimhanesi, S., Ugbodu, F. & Iyore, N. (2024). Heliyon Data-Driven Intelligent Modeling, Optimization, and global sensitivity analysis of a xanthan gum biosynthesis process. *Heliyon*, 10(3), e25432. https://doi.org/10.1016/j.heliyon.2024.e25432

Nouri, A., Vaziri, H., Belhaj, H., & Islam, R. (2006). Sand-production prediction, a new set of criteria for modeling based on large-scale transient experiments and numerical investigation. *SPE Journal, 11*(2), 26–29.

Papamichos, E., & Furui, K. (2019). Analytical models for sand onset under field conditions. *Journal of Petroleum Science and Engineering, 172*, 171–189.

Shabdirova, A., Minh, N. H., & Zhao, Y. (2022). Role of plastic zone porosity and permeability in sand production in weak sandstone reservoirs. *Underground Space (China)*. https://doi.org/10.1016/j.undsp.2021.10.005

Shabdirova, A., Kozhagulova, A., Samenov, Y., & others. (2024). Sand production prediction with machine learning using input variables from geological and operational conditions in the Karazhanbas oilfield, Kazakhstan. *Natural Resources Research, 33*, 2789–2805. https://doi.org/10.1007/s11053-024-10389-3

Sheridan, R. P., Wang, W. M., Liaw, A., Mu, J. & Gifford, E. M. (2016). Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationship. *Journal of Chemical Information and Modeling*, 56(12), 2353-2360. https://doi.org/10.1021/acs.jcim.6b00591

Skjaerstein, A., Tronvoll, J., Santarelli, F. J., & Joranson, H. (1997). Effect of water breakthrough on sand production, experimental and field evidence. In *Proceedings - SPE Annual Technical Conference and Exhibition Pi*, 565–575.

Song, J., Li, Y., Liu, S., Xiong, Y., Pang, W., He, Y. & Mu, Y. (2022). Comparison of Machine Learning Algorithms for Sand Production Prediction: An Example for a Gas-Hydrate-Bearing Sand Case. *Energies 2022*, 15(18), 6509. https://doi.org/10.3390/en15186509

Speiser, J. L., Miller, M. E., Tooze, J. & Ip, E. (2019). A Comparison of Random Forest Variable Selection Methods for Classification Prediction Modeling. *Expert Systems with Applications*, 134, 93–101. https://doi.org/10.1016/j.eswa.2019.05.028

Tiab, D. & Donaldson, E. C. (2004). Petrophysics: Theory and Practice of Measuring Reservoir Rock and Fluid Transport Properties, *Gulf Professional Publishing, Elsevier* (2nd. Ed., pp. 554-670)

Tixier, M.P. (1949) Evaluation of Permeability from Log Resistivity Gradients. Oil and Gas Journal, 48, 113-122.

Wang, Y., & Dusseault, M. B. (2010). Sand production potential near inclined perforated wellbores. In *47th Annual Technical Meeting of the Petroleum Society*. https://doi.org/10.2118/9670

Weingarten, J. S., & Perkins, T. K. (2007). Prediction of sand production in gas wells, methods and Gulf of Mexico case studies. *Journal of Petroleum Technology, 47*(7), 596–600.

Wu, B., Choi, S. K., Denke, R., Barton, T., Viswanathan, C., Lim, S., Zamberi, M., & Shaffee, S. (2016). A new and practical model for amount and rate of sand production. *Offshore Technology Conference.*