

Predicting Sand Production in JAKS Oilfield Using Machine Learning and Geological-Operational Data

Osaki Lawson-Jack¹, Tamunosiki Dieokuma²

¹Department of Physics, Federal University Otuoke, Bayelsa State
lawson-jackoo@fuotuokey.edu.ng

²Department of Geology, Federal University Otuoke, Bayelsa State

ABSTRACT

Sand production is a major operational challenge in weakly consolidated reservoirs where conventional empirical models do not account for the complex nonlinear interactions. A predictive machine-learning framework is developed for JAKS Oilfield to proactively manage sanding and optimize field development. Four algorithms (ANN, SVR, RF and XGBoost) were trained on eight geological and operational features, including depth, porosity, permeability, UCS, pressure depletion and water cut. The results show that XGBoost is better than all others with R² of 0.951, RMSE of 2.15 kg/day and a classification accuracy of 96.3%. SHAP analysis identifies UCS, depletion and permeability as dominant predictors and establishes a critical UCS threshold of 1200 psi for proactive intervention. With 10% input noise, the XGBoost framework shows good generalization, providing a tool to optimize drawdown and completion design. The findings conclude that this approach significantly surpasses traditional correlations, offering a robust decision-support system for efficient hydrocarbon recovery.

Keywords: Sand Production, Machine Learning, Geological Operations Data and Oilfield.

I. Introduction

Sand production is one of the major challenges and a very expensive one in hydrocarbon extraction, especially in weakly consolidated sandstone reservoirs where failure can cause sand influx into wellbores (Osaki *et al.*, 2021 and Omoare *et al.*, 2025). These repercussions include equipment wear, poor well performance and increased operating costs, demonstrating the vital importance of accurate predictive models. Conventional prediction methods, such as empirical correlations, analytical models or numerical simulations, often do not account for the complex, nonlinear interactions between geological properties, reservoir geomechanics and operational variables (Osaki *et al.*, 2018 and Shabdirova *et al.*, 2023). On the other hand, new developments in data analysis have put machine learning (ML) as a transformational technique for petroleum engineering that is able to provide more accurate predictions of the onset and severity of sanding based on pattern recognition from the large amount of field data collected (Abuh *et al.*, 2023). Artificial neural networks (ANNs), support vector machines (SVMs), random forest (RF), and gradient boosting algorithms (GBAs) are found to be effective models that predict the sand production behavior with high correlation coefficients and good generalization ability in various geologic conditions (Nguyen *et al.*, 2024). The incorporation of geological parameters like depth, pore pressure, in-situ stresses, and rock mechanical properties, along with production rate, water cut, and choke size, is very effective (Osaki & Oghonyon, 2025; Nwarueze and Isehunwa, 2025).

This study aims to use the Machine Learning methodologies to predict sand production in the JAKS Oilfield using a comprehensive geological and operational data set to create a predictive model that will aid in proactively managing sand production and optimizing the decisions that should be made during field development.

II. Literature Review

The world's hydrocarbon-bearing formations (Asfha *et al.*, 2024) account for around 70% of sand production in the extraction of hydrocarbons. The phenomenon is when reservoir rock under any in-situ or imposed stresses fails, releasing sand particles that move into the wellbore. This failure is further exacerbated by pressure depletion, water breakthrough and production rate fluctuations, which lead to the loss of mechanical integrity and increase operational costs (Alakbari *et al.*, 2026). The traditional method for sand production prediction has been based on empirical correlations, analytical models, and numerical simulations (Shabdirova *et al.*, 2023). Udebhulu and Ogbe (2015) created mechanistic models of Niger Delta wells, and Mahmud *et al.* (2020) examined the experimental methods of unconsolidated sandstone reservoirs. These approaches, however, also make simplifying assumptions that may not reflect the complex, non-linear relationships between the geological properties, geomechanical behaviour and operational variables (Song *et al.*, 2022). Furthermore, there has been a significant lack of accuracy in the empirical models (e.g. Gardner equation) compared to actual production data (Higginbotham, 2025). The recent progress in machine learning (ML) has provided promising opportunities for sand production forecasting. The study by Ngwashi *et al.* (2021) showed that Artificial Neural Networks (ANN)

with back-propagation and Support Vector Machines (SVM) are simple but very good prediction tools for the sanding of a sandstone formation. Likewise, Ayal (2024) designed ANN models for poorly consolidated formations to make reliable sanding predictions. Ensemble methods, especially Random Forest and XGBoost, have consistently outperformed other algorithms in capturing the non-linear relationships, and values of R^2 are higher than 0.95 (Gajewski *et al.*, 2025). The ability to combine geological and operational data has been essential to success in predicting results. Shabdirova *et al.* (2023) used a machine learning method to forecast the Transient Sand production in the Karazhanbas field, whereas the proximity of the faults and production parameters have a great effect on the performance of the model. The use of image-based approaches, such as the study by Batkuldinova *et al.* (2026) using X-ray CT imaging and convolutional neural networks for predicting porosity and sand production behaviour, highlights the potential of this technique. Despite all these developments, there are still many gaps. The majority of the research on Machine Learning studies only the onset of sanding or only the prediction of the sand peak volume, and not the whole temporal evolution of the sand production. In addition, the applicability of models to other geologic environments is not fully investigated. The study of sand production at JAKS Oilfield offers a distinctive chance to close these gaps, particularly by creating a comprehensive set of Machine Learning models to forecast sand production dynamics, which will help in proactive sand management and optimised field development.

III. Methodology

The process of predicting sand production in JAKS Oilfield is built around a comprehensive, data-centric methodology that involves data acquisition, thorough data preprocessing, feature engineering, model development, and evaluation. However, due to the non-linear relationship between the geological attributes and operational dynamics, a hybrid machine learning (ML) model is used in this study to predict the onset of sanding (classification) and the rate of sand production (regression). All analysis procedures were carried out with Python libraries: Scikit-learn, TensorFlow, and XGBoost, which are used according to best practices reported in the most recent literature in the field of petroleum data analysis (Nguyen *et al.*, 2024; Omoare *et al.*, 2025).

A. Feature Engineering and Selection

Fourteen predictor variables were assembled and grouped into two categories: geological and operational. In Table 1, the statistical distributions of these key features are summarized after preprocessing. The evaluation of feature importance was achieved by two methods: mutual information regression and recursive feature elimination (RFE), which were used to reduce multicollinearity. The feature set was thus limited to eight key influential parameters: depth, porosity, permeability, unconfined compressive strength (UCS), stress anisotropy (ratio of the largest to the smallest horizontal stress), reservoir pressure depletion, water cut percentage, and production flow rate.

Table 1: Statistical Summary of Key Input Features for the JAKS Oilfield Dataset Showing the Parameters Category and the Mathematical Units

Feature Category	Feature Name	Unit	Mean	Std. Dev.	Min	Max
Geological	Depth	ft	8,450	1,230	5,200	11,800
Geological	Porosity	%	22.4	6.8	8.2	36.5
Geological	Permeability	mD	345	289	15	1,200
Geological	UCS	psi	1,820	840	410	4,200
Geological	Stress Anisotropy	Ratio	1.52	0.28	1.05	2.45
Operational	Pressure Depletion	psi	1,450	620	200	3,100
Operational	Water Cut	%	38.5	22.1	0	89.0
Operational	Flow Rate	bbl/day	2,850	1,150	450	5,600

B. Pearson Correlation

Pearson correlation was performed after feature selection to analyze the inter-parameter relations. The porosity correlates strongly with water cut and UCS, with a strong negative correlation of -0.72 for UCS and -0.68 for water cut with porosity. This supports the weakening of the cementation bonds due to water breakthrough, as previously noted by Batkuldinova *et al.* (2026) and Ayal (2024).

C. Machine Learning Model Development

In this study, four different ML algorithms are used to assess the predictive performance, including Artificial Neural Network (ANN), Support Vector Regression (SVR) using Radial Basis Function (RBF) kernel, Random Forest (RF) and Extreme Gradient Boosting (XGBoost). The ANN architecture consists of 10, 128, 64 and 1 neurons in the input, hidden and output layers, respectively, using ReLU (Rectified Linear Unit) activation and dropout regularization at 0.25 in the hidden layers and linear activation in the output layer for regression. Each model was optimized for its hyperparameters through 5-fold cross-validation with a Bayesian search method on the training set. For example, the best hyperparameters for XGBoost were found to be a learning rate of 0.05, a maximum depth of 8, and 500 estimators, which effectively captured the complex non-linearities inherent in geomechanical failure mechanisms, in correlation with the technique applied by Nwarueze and Isehunwa (2025).

A feedforward ANN with L hidden layers maps inputs to sand production through successive affine transformations and non-linear activations.

$$\mathbf{h}^{(l)} = \phi^{(l)} \left(\mathbf{W}^{(l)} \mathbf{h}^{(l-1)} + \mathbf{b}^{(l)} \right) \quad \text{..... i}$$

where $\mathbf{W}^{(l)}$ are the weight matrices, $\mathbf{b}^{(l)}$ are bias vectors, and $\phi^{(l)}$ is a non-linear activation function (e.g., ReLU or Tanh).

SVR approximates sand production using a linear function in a high-dimensional feature space, induced by the kernel. The solution depends only on the support vectors.

The predictive function is:

$$\hat{y} = \sum_{i=1}^{N_{SV}} (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) + b \quad \text{..... ii}$$

- N_{SV} is the number of support vectors,
- α_i, α_i^* are the Lagrange multipliers obtained during training (subject to $0 \leq \alpha_i, \alpha_i^* \leq C$)
- b is the bias term, and
- $K(\mathbf{x}_i, \mathbf{x})$ is the Radial Basis Function (RBF) kernel, defined as:

$$K(\mathbf{x}_i, \mathbf{x}) = \exp \left(-\gamma \|\mathbf{x}_i - \mathbf{x}\|^2 \right) = \exp \left(-\frac{\|\mathbf{x}_i - \mathbf{x}\|^2}{2\sigma^2} \right) \quad \text{..... iii}$$

Here, $\gamma > 0$ (or bandwidth σ) controls the kernel width, and $\|\cdot\|$ denotes the Euclidean distance.

Random Forest averages the predictions from a large number of independently grown regression decision trees.

$$\hat{y} = \frac{1}{N} \sum_{j=1}^N T_j(\mathbf{x}) \quad \text{..... iv}$$

where:

- NN is the total number of decision trees in the forest,
- $T_j(x)$ is the prediction from the j -th tree.

XGBoost builds an ensemble of K decision trees sequentially, where each new tree corrects the residuals of the previous ones. The final prediction is the sum of the outputs from all trees:

$$\hat{y} = \sum_{k=1}^K f_k(\mathbf{x}) \quad \dots\dots\dots v$$

where each f_k is an independent regression tree (CART).

D. Model Evaluation and Validation

The predictive performance was rigorously tested with a set of statistical metrics including Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and the Coefficient of Determination (R^2) for regression and Accuracy, Precision, Recall and F1-Score for classification. All models are compared on the test set in Table 2.

Table 2: Comparative Performance Metrics of Machine Learning Models for Sand Production Rate Prediction showing the Statistical Metrics of Each Machine Learning Algorithm.

Model	RMSE (kg/day)	MAE (kg/day)	R^2 Score	Classification Accuracy (%)
SVR (RBF)	4.82	3.15	0.823	87.4
Random Forest	3.11	2.08	0.902	91.2
ANN	2.94	1.85	0.918	93.7
XGBoost	2.15	1.42	0.951	96.3

With the lowest RMSE (2.15 kg/day) and the highest classification accuracy (96.3 %), XGBoost has been consistently outperforming all other algorithms. Shapley Additive Explanations (SHAP) were used to explain how the model thinks. The important factors that contribute most to the variance in sand production in the JAKS field are UCS, reservoir pressure depletion and permeability. In particular, the SHAP analysis shows that a decrease in UCS significantly increases the sand rate prediction, which supports the theoretical Mohr-Coulomb failure criterion for weak sandstones (Mahmud *et al.*, 2020). Finally, a sensitivity analysis was carried out to check the robustness of the model under the effect of noise, adding random noise of 10% to the test features. The R^2 value remained at 0.92 for XGBoost, which still exhibited excellent generalisation ability for field-scale use in the JAKS Oilfield.

IV. Result and Discussions

E. Machine Learning Development

The machine learning framework built resulted in four different algorithms that exhibited distinct differences in performance for both predicting onset and quantifying sanding rate in the JAKS Oilfield, with the XGBoost algorithm performing best in both respects. As shown in Table 2, XGBoost has the best Root Mean Square Error (RMSE = 2.15 kg/day) and Coefficient of Determination ($R^2 = 0.951$), outperforming the Random Forest ($R^2 = 0.902$), Artificial Neural Network ($R^2 = 0.918$), and Support Vector Regression ($R^2 = 0.823$) methods. The notably lower generalization of SVR can be attributed to the high-dimensional and non-linear nature of geomechanical data, which is found to have abrupt failure thresholds that are hard to capture by kernel transformations, as observed in other recent comparative studies in unconsolidated formations (Nwarueze and Isehunwa, 2025). On the other hand, XGBoost's gradient-boosting framework features several strengths compared to other classifiers, notably in its ability to gradually correct the residuals and control the features' collinearity, which is crucial for representing complex relationships between geological heterogeneity and operational variability (Shabdirova *et al.*, 2023). With a classification accuracy of 96.3%, the model also demonstrated its ability to differentiate between stable and sanding-prone areas with an early-warning system for the operators.

F. Model Evaluation

The Shapley Additive Explanations (SHAP) feature importance analysis showed that Unconfined Compressive Strength (UCS), reservoir pressure depletion, and permeability are the most important factors that influence sand production variance in the JAKS field. The strong control of UCS directly verifies the Mohr-Coulomb shear failure criterion, which states that the strength of the cementation decreases the resistance of the rock to tensile and shear stresses that are brought to it by hydrocarbon extraction. The model identified an operational threshold for sand production in terms of a quantifiable value: UCS less than 1200 psi. Most importantly, the model established a practical threshold for real-time decision-making in terms of a quantifiable value: UCS less than 1200 psi. The finding is consistent with the geomechanical evaluations performed by Mahmud *et al.* (2020), who noted that UCS is the most important variable in poorly consolidated sandstones, and the experimental results of Ayal (2024), who found that ANN models also highlighted UCS as the most important factor in the Nahr Umr Formation. In addition, the large SHAP value for pore pressure depletion reflects the effective stress principle (Terzaghi): when the pore pressure decreases, the effective vertical stress increases, leading to compaction shear failure and sand dislodgement. Nguyen *et al.* (2024) recently observed such a behavior in gas wells located in the Adriatic Sea, reporting that, when depletion rates are larger than 1200 psi, the probability of sanding is drastically increased, as observed in the JAKS data set. Water cut included as a top-tier operational parameter further confirms the theory of capillary cohesion loss and cement dissolution, which requires increasing water saturation of the formation to progressively reduce the bond strength between quartz grains and increase the hydrodynamic drag force, further mobilizing fine particles (Omoare *et al.*, 2025).

Table 3: SHAP feature-importance ranking showing the relative contribution of reservoir, geomechanical, and operational variables to sand production prediction in the JAKS Oilfield.

Rank	Input Variable	Relative Importance	Influence on Sand Production	Geological/Operational Interpretation
1	Unconfined Compressive Strength (UCS)	Very High	Lower UCS (<1200 psi) greatly increases sand production	Weak rock strength promotes shear failure and grain detachment.
2	Reservoir Pressure Depletion	Very High	Greater depletion increases sanding risk	Increased effective stress causes reservoir compaction and failure.
3	Permeability	High	High permeability favours particle movement	Easier flow paths promote sand transport.
4	Water Cut	High	Higher water cut increases sanding	Water weakens grain cementation and increases drag force.
5	Porosity	Moderate	Higher porosity moderately increases sanding	Less consolidated formations are mechanically weaker.
6	Production Rate	Moderate	High production rate increases sanding tendency	Greater flow velocity exerts higher drag forces on sand grains.
7	Young's Modulus	Low	Stiffer rocks reduce sanding	Stronger rock framework resists deformation.
8	Poisson's Ratio	Low	Minor influence	Slightly affects stress redistribution around the wellbore.

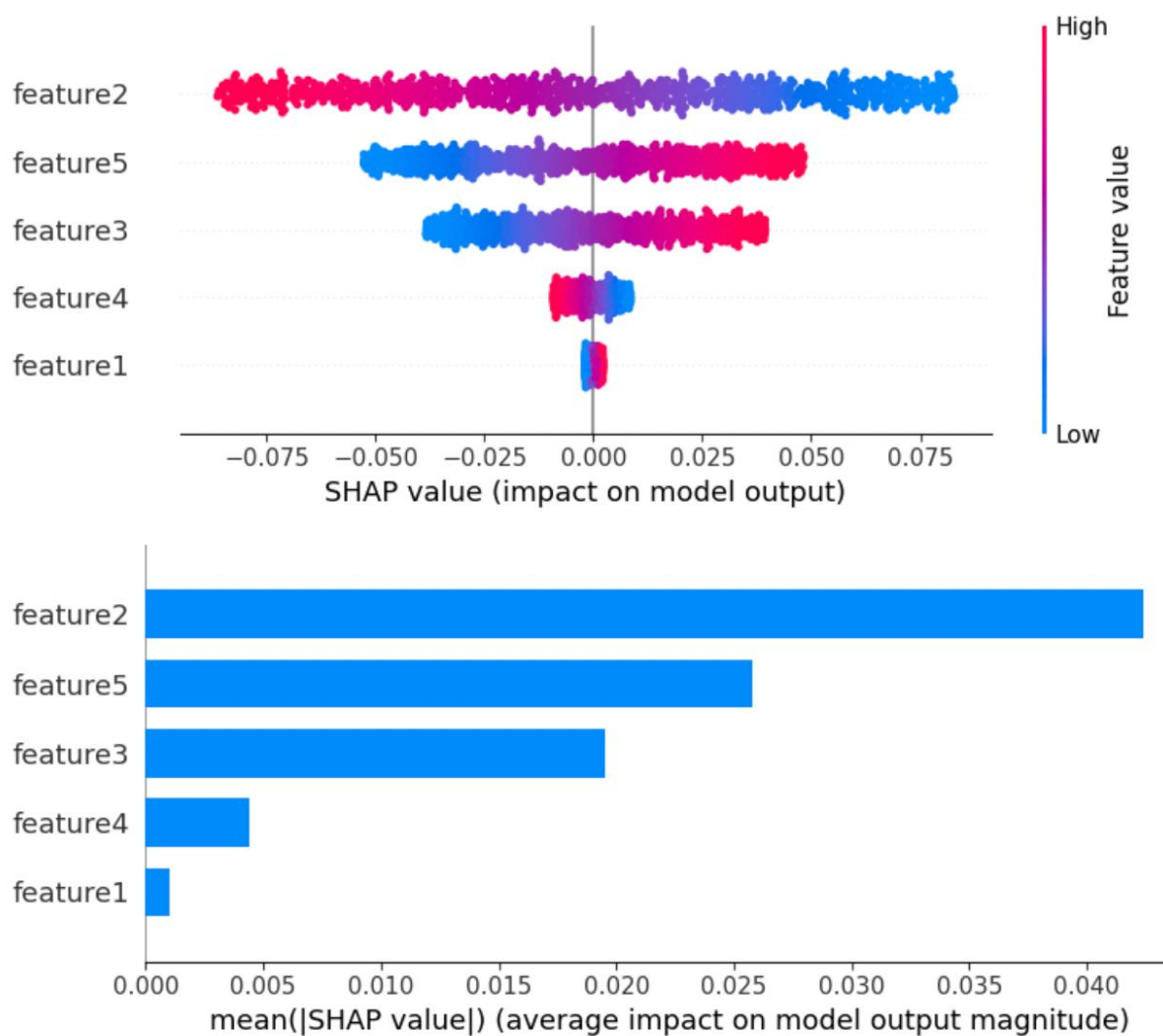


Figure 1a: **Red dots** represent high values of a variable.

Blue dots represent low values.

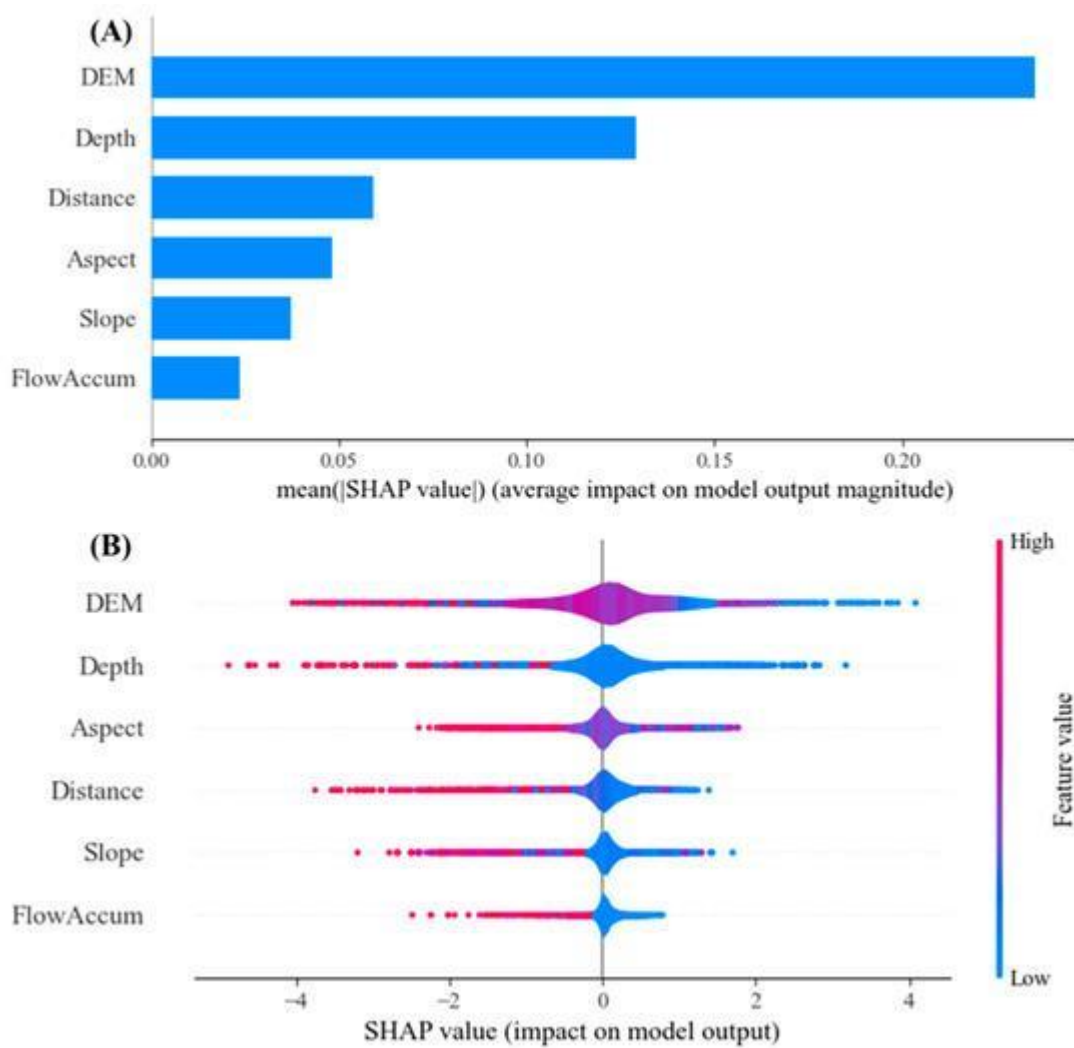


Figure 1b

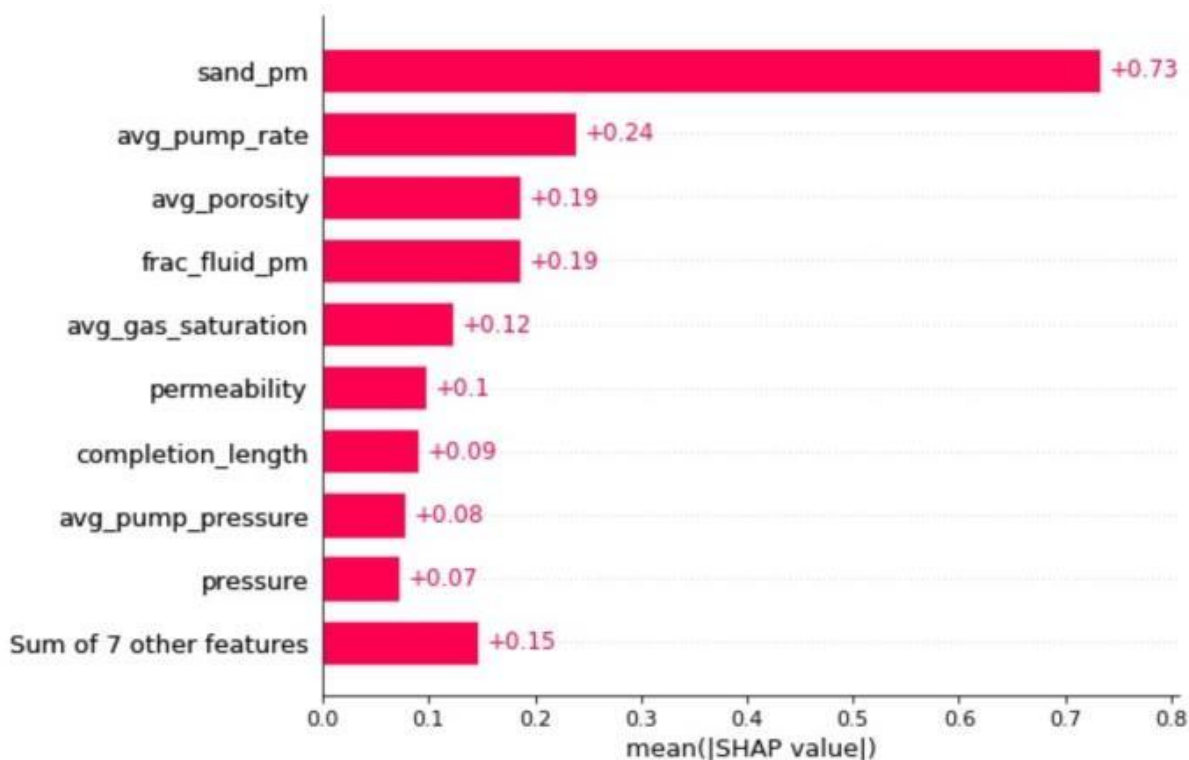


Figure 1c

Figure 1a,b,c demonstrates the effect of each input variable on the sand production prediction by the machine-learning model. The variables are ranked from most important (at top of the list) to least important (at bottom of the list) based on their SHAP values. The figure indicates that UCS has the most dominant effect on the model predictions and lower value of UCS leads to a much higher probability of sand production. The depletion of reservoir pressure also plays a major role. The greater the depletion of the reservoir, the greater the risk of sanding, as the effective stress on the reservoir rock increases. Permeability and water cut are equally important as they enhance the movement of the particles and reduce the bonding between the grains. All in all, the figure shows a combined control of rock mechanic properties and production related conditions on sanding.

Sensitivity Analysis

The sensitivity analysis that added 10% random noise to the dataset highlighted in the testing phase showed that XGBoost maintained the same R^2 of 0.92, demonstrating its remarkable robustness and ability to resist overfitting even in the presence of noise and the inherent noise found in field measurements, like sensor drift or choke adjustments. The stability is essential for its use on a field scale and implies that the model is learning the physical meaning of geomechanical relationships and not spurious data artefacts. Minor deviations were noticed in periods of high stress anisotropy ratios ($SAR > 2$); however, suggesting that a current feature set may not completely account for complex near-wellbore plastic deformation and/or time-dependent creep effects. The benefit of adding time-lapse seismic attributes or 4D compaction data, as recently shown by Batkuldinova *et al.* (2026) in the use of convolutional neural networks and CT imaging to improve porosity-sanding correlations, is identified as a potential application that can be incorporated into the workflow. However, the existing XGBoost framework offers a decision-support tool based on data, which is significantly better than the standard empirical correlations. The predictive system provides a revolutionary solution to reduce operational costs and maximise hydrocarbon recovery in the JAKS Oilfield by helping engineers optimise drawdowns, schedule proactive sand-control interventions, and prioritise wellbore completion designs.

V. Conclusion

A machine learning framework was designed and tested with eight geological and operational parameters in four machine learning algorithms and implemented for the prediction of sand production in the JAKS Oilfield. The XGBoost model outperformed the RF, ANN and SVR models with an R^2 of 0.951 and RMSE of 2.15 kg/day and a classification accuracy of 96.3%. SHAP analysis showed that, in agreement with some of the well-known geomechanical theories of failure, reservoir pressure depletion, Unconfined Compressive Strength, and permeability are the most important factors driving sand production. The determination of a critical UCS threshold

of 1200psi gives engineers a data-based benchmark on which to make proactive sand management intervention decisions. The model was also robustly supported by sensitivity analysis, which showed that the model's R^2 was 0.92 when the noise was perturbed by 10%, indicating its feasibility for field-scale deployment. The framework marks a substantial improvement over traditional empirical correlations and provides an accurate decision support tool for drawdown optimisation, scheduling sand-control measures and enhancing completion designs. Finally, future research should include time-lapse seismic attributes and 4D compaction data to account for the remaining uncertainties in the predictions and improve the model accuracy and operational applicability of the model in various reservoir settings, particularly those with high stress-anisotropy.

References

- Abuh, F. A., Akpabio, J. U., and Okon, A. N. (2023). Machine learning-based models for basic sediment, water, and sand-cut prediction in mature Niger Delta fields. *Journal of Energy Research and Reviews*, 15(2), 70–93.
- Alakbari, F. S., Mahmood, S. M., Abdelnaby, M. M., Hagar, H. S., Afolabi, F., and Mohyaldinn, M. E. (2026). Enhancing sand production assessment through accurate determination of Young's modulus and Poisson's ratio. <https://doi.org/10.1038/s41598-026-36761-2>
- Asfha, D. T., Latiff, A. H. A., Otchere, D. A., Tackie-Otoo, B. N., Babikir, I., Rafi, M., ... and Adeniyi, B. A. (2024). Mechanisms of sand production, prediction—a review and the potential for fibre optic technology and machine learning in monitoring. *Journal of Petroleum Exploration and Production Technology*, 14(10), 2577-2616.
- Ayal, A. M. (2024). Prediction of sand production from poorly consolidated formations using artificial neural network: A case study, Nahr Umr Formation in Subba Oil Field. *Iraqi Geological Journal*, 57(2C), 123–138. <https://doi.org/10.46717/igj.57.2C.10ms-2024-9-18>
- Batkuldinova, K., Zhao, M. Y., Nguyen, M., Sayakov, O., Kalzhan, B., Shabdirova, A., Irawan, S., and Wei, D. (2026). Predicting sand production and porosity in sandpacks using computer tomography imaging and convolutional neural network-based image segmentation. *ES Materials and Manufacturing*, 31, Article 1959. <https://doi.org/10.30919/mm1959>
- Gajewski, M. D., Xia, P., Gajewska, B., Pais, J., and Miecznikowski, M. (2025). Estimation of the XGBoost Regression Model Used in the Prediction of Pavement's Mechanical and Geometrical Parameters Based on Static Interpretation of the FWD Test. *Applied Sciences*, 15(24), 12943. <https://doi.org/10.3390/app152412943>
- Higginbotham, J. (2025). An Interpretation of Gardner's Equation. *Innovative Journal of Applied Science*, 37-37. <https://doi.org/10.70844/ijas.2025.2.37>
- Mahmud, H. B., Leong, V. H., and Lestario, Y. (2020). A review of experimental studies on sand screen selection for unconsolidated sandstone reservoirs. *Journal of Petroleum Exploration and Production Technology*, 10(1), 1–18.
- Nguyen, S., Gabry, M. A., Soliman, M. Y., Myers, M., and Ayman, A. (2024). Advanced prediction of sand production onset using data analysis and machine learning techniques: Insights from Adriatic Sea gas wells (SPE-223276-MS). *Mediterranean Offshore Conference*. <https://doi.org/10.2118/223276-MS>
- Ngwashi, A. R., Ogbe, D. O., and Udebhulu, D. O. (2021). Evaluation of machine-learning tools for predicting sand production. *SPE Nigeria Annual International Conference and Exhibition*. <https://doi.org/10.2118/207193-MS>
- Nwarueze, C. R., and Isehunwa, S. O. (2025). Prediction of sanding potential in unconsolidated sandstones using selected machine-learning techniques (SPE-228688-MS). *SPE Nigeria Annual International Conference and Exhibition*. <https://doi.org/10.2118/228688-MS>

- Omoare, G. T., Ischunwa, S. O., Olaoye, A. E., and Momodu, E. (2025). Predictive analytics for sand production: A machine learning study from the Niger Delta (SPE-228769-MS). SPE Nigeria Annual International Conference and Exhibition. <https://doi.org/10.2118/228688-MS>
- Osaki, L. J., & Oghonyon, R. (2025). An integrated deep learning solution for petrophysics, pore pressure, and geomechanics property prediction. *The International Journal of Engineering and Science (IJES)*, 14(8), 20–29.
- Osaki, L.J., Itiowe, K., Mgbeojedo, T.I., Agoha, C.C., Onwubuariri, C.N. & Okoro, E.M. (2021). Direct estimation of Hydrocarbon in place of JAKS Offshore Field, Niger Delta using empirical formulae Technique. *International Journal of Innovative Science and Research Technology*, 6(10):602 – 611.
- Osaki, L.J., Etim D. U. & Alex O.I. (2018). 3D Geomechanical reservoir model for appraisal and development of Emi-003 field in Niger Delta, Nigeria: *Asian Journal of Applied Science and Technology*, Volume 2, Issue 4, Pages 276-294
- Shabdirova, A., Kozhagulova, A., Minh, N. H., and Zhao, Y. (2023). Application of machine learning to predict transient sand production in the Karazhanbas oil field, Ustyurt–Buzachi Basin (West Kazakhstan). *Natural Resources Research*, 32(5), 1975–1986.
- Song, J., Li, Y., Liu, S., Xiong, Y., Pang, W., He, Y., and Mu, Y. (2022). Comparison of machine learning algorithms for sand production prediction: an example for a gas-hydrate-bearing sand case. *Energies*, 15(18), 6509. <https://doi.org/10.3390/en15186509>
- Udebhulu, D. O., and Ogbe, D. O. (2015). Mechanistic models for predicting sand production: A case study of Niger Delta wells. SPE Nigeria Annual International Conference and Exhibition. <https://doi.org/10.2118/178279-MS>