



Original Paper

# Sand Production Prediction with Machine Learning using Input Variables from Geological and Operational Conditions in the Karazhanbas Oilfield, Kazakhstan

Ainash Shabdirova <sup>1,2,4</sup> Ashirgul Kozhagulova,<sup>2</sup> Yernazar Samenov,<sup>2</sup> Nguyen Minh,<sup>3</sup> and Yong Zhao<sup>1</sup>

Received 1 March 2024; accepted 17 July 2024  
Published online: 9 August 2024

This paper describes a comprehensive approach to predict sand production in the Karazhanbas oilfield using machine learning (ML) techniques. By analyzing data from 2000 wells, the research uncovered the complex dynamics of sand production and emphasized the critical need for accurately predicting the peak sand mass and its occurrence time. ML techniques can have a significant impact on prediction of sand production and on the optimization of oilfield operation, which can be improved with the combined use of enriched training data and domain-specific knowledge. The research underscored the influence of geological factors, especially fault proximity, on prediction accuracy. Domain and field knowledge is needed to formulate different production scenarios for prediction purposes such that the relevant data can be selected for the training of ML models. Moreover, new metrics are needed to evaluate model performance as the applied method is tailored for different operational strategies. As the peak sand mass is considered a pivotal event in field operation, new metrics in terms of peak prediction accuracy and peak time prediction accuracy were introduced to evaluate the performance of ML models. A suite of ML algorithms was employed in the study, which demonstrated notable accuracy in the classification of sand-producing wells.

**KEY WORDS:** Sand production, machine learning, prediction accuracy.

## INTRODUCTION

Sand production is a critical issue, especially prevalent in weaker, unconsolidated formations. Its significance in the industry is underscored by the multifaceted challenges it presents, ranging from equipment erosion and increased maintenance

requirements to environmental concerns. The dynamics of sand production exhibits varied patterns in the field, often characterized initially by a burst, followed by potential stabilization or decline. This transient nature of sand production hinges on key parameters such as the peak production rate and timing of its occurrence. Traditionally, the industry's approach to managing sand production has revolved around reactive measures post-occurrence. However, recent advancements in predictive analytics, particularly the application of machine learning (ML) algorithms, have opened new frontiers in proactive management.

<sup>1</sup>School of Engineering and Digital Sciences, Nazarbayev University, Astana, Kazakhstan.

<sup>2</sup>Oil and Gas Faculty, Atyrau University of Oil and Gas, Atyrau, Kazakhstan.

<sup>3</sup>Fulbright University Vietnam, Ho Chi Minh City, Vietnam.

<sup>4</sup>To whom correspondence should be addressed; e-mail: ainash.shabdirova@nu.edu.kz

Various sand production patterns can be observed in the field, as noted by Fjaer et al. (2008): Initially, when a well is started, there is a surge in sand production due to the release of perforation debris with reservoir fluid. At certain flow rates, equilibrium is established, leading to transient sand production. The key parameters of interest in transient sand production are the peak value and declining rate. Accurate prediction of these parameters is crucial for planning sand disposal measures. For instance, if the peak value is not critical and the declining rate is high, indicating a small amount of sand over a short period, there may be no need for long-term sand storage. Conversely, a high peak value with a low declining rate necessitates preparation for significant sand production over an extended period, requiring specific sand disposal sites. Therefore, effective prediction of peak sand production and declining rates forms the foundation of a cost-effective sand management strategy.

Comprehensive research has been conducted to predict sanding onset, and as a result, there are many analytical, numerical, and empirical models which can predict critical production conditions triggering sand production to a certain extent (Morita et al. 1989; Kessler et al., 1993; Weingarten and Perkins, 2007; Wang and Dusseault, 2010; Han et al., 2011; Al-Shaabi et al., 2013; Fuh and Morita, 2013; Wu et al. 2016; Papamichos and Furui, 2019). On the other hand, predicting not only sanding onset but also its volume is crucial for sand management strategies justification. However, sand volume prediction models are susceptible to specific parameters and assumptions and therefore can be only applied to certain field conditions. Moreover, sand volume prediction is particularly essential because it affects the sand control solutions planned for a specific well.

The principle of analytical sand volume prediction models is to establish a mathematical formulation for the constitutive material behavior and sand detachment mechanism (Geilikman et al., 1994; Al-Shaabi et al., 2013; Papamichos and Furui 2013; Gholami et al. 2016; Hayavi and Abdideh 2017; Papamichos and Furui 2019; Shabdirova et al., 2019). Hole stability analysis is then conducted by introducing failure criteria to a constitutive model. This is followed by the analytical expression for the sand detachment process. Analytical models are fast and easy to use; however, they mostly simulate very simplified conditions. Because any model employs simplifying assumptions, analytical models can be

used where a quick and reliable sand prediction tool is required.

More complex conditions can be considered if the numerical prediction methods are utilized (Yi, 2003; Nouri et al. 2006b; Kim et al., 2011; Chen, 2012; Cui et al. 2016; Han and Cundall 2017; Li et al., 2018; Wang et al., 2018, 2019a, 2019b; Garolera et al., 2019; Khamitov et al., 2022). Numerical models can be grouped under continuum and discontinuum methods. In the continuum method, rock is treated as a continuous material, and constitutive laws are employed when describing its behavior. The main disadvantage of continuum models is that they cannot describe local discontinuities as particle detachment. In this case, discontinuum methods are used. Three-dimensional discrete element modeling (DEM) can capture the motion of individual grains and therefore provide information on sanding mechanism at micro-scale. Considering each individual particle makes DEM computationally expensive and time-consuming, which prevents using DEM for large-scale problems.

Laboratory experiments are very useful for studying the mechanisms of sand production, although their results are strongly influenced by boundary conditions and they are not always straightforward to apply to field-scale problems (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). Experimental results are mostly used to calibrate and validate analytical and numerical sand production models.

Despite advancements in analytical, numerical, and experimental techniques for predicting sand production, these methodologies are often limited by their reliance on simplifying assumptions, computational intensity, and the challenges of scaling from laboratory conditions to field applications. ML offers a promising alternative by leveraging vast amounts of data to identify complex patterns and relationships that traditional models may overlook. By integrating ML with existing predictive models, the industry can achieve a more comprehensive and robust understanding of sand production dynamics, leading to better-informed decision-making and reduced operational risks.

The prediction of sanding onset can be viewed as a classification task with two outcomes—whether sanding will or will not occur under given conditions. Several studies have successfully applied ML algorithms to predict sanding onset and analyze the

significance of input variables. Khamehchi et al. (2014) presented an analysis of sand production onset in 23 field datasets from the North Adriatic Sea, utilizing 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. The algorithm, evaluated using receiver operating characteristics (ROC) curve, demonstrated high predictive capability. Ketmalee and Bandyopadhyay (2018) employed 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 back-propagation and SVM in predicting sanding onset in the Niger Delta region, reporting 80% and 100% accuracy, respectively. Abdelghany et al. (2022) used a probabilistic neural network to infer reservoir properties and predict sand production onset, reporting improved accuracy compared to conventional models. Song et al. (2022) applied four ML algorithms to predict sand production from gas-hydrate sediments, recommending XGBoost for early-stage prediction.

While there are significant researches on using ML for classifying the onset of sand production, less focus has been on quantitatively predicting the amount of sand produced. Most studies aimed to guide the decision on whether to implement sand control measures during the well completion phase. However, the actual amount of sand produced is crucial for planning surface facilities and cost-effective management, indicating a need for more research in this area.

This manuscript builds upon our previous work published (Shabdirova et al., 2023), whereby we explored the potential of ML algorithms in predicting transient sand production in the Karazhanbas oilfield in Western Kazakhstan. Our initial study, based on data from 43 wells, established a foundational understanding of the applicability of the Random Forest (RF) algorithm for predicting sand production behaviors. Building on these findings, the current research, which now encompasses data from 2000 wells, placed greater emphasis on refining data preparation methodologies rather than exclusively focusing on ML techniques. This strategic shift allowed for a more thorough investigation into the intricacies of data preprocessing, aiming to improve the quality of input data. This expansion is instru-

mental in facilitating a more comprehensive analysis, enhancing predictive accuracy, and enabling the exploration of more complex aspects of sand production prediction.

This paper is structured as follows: The Oilfield Information section gives background information on the oilfield. The Methodology section outlines an approach to sand production prediction, incorporating data collection, ML algorithms, model training, and validation, as well as input variable analysis. The Results and Discussion section of the manuscript assesses various operational and geological factors influencing sand production in the Karazhanbas oilfield. It demonstrates the importance of data diversity and quality in improving prediction accuracy, highlights the significant impact of fault proximity on predictions, and introduces novel metrics for evaluating model performance, especially regarding peak sand production events. The manuscript concludes with a summary of the main findings, emphasizing their implications for the industry and suggesting directions for future research.

## Oilfield Information

The Karazhanbas oilfield, located in the Ustyurt-Buzachi sedimentary basin in Western Kazakhstan, is characterized by a gentle, elongated anticlinal fold, flanked by western and eastern arches (Fig. 1). This field's structure is divided by a series of faults. Its reservoirs are found within the Lower Triassic, Middle Jurassic, and Lower Cretaceous strata. Despite the region's tectonic changes, these formations remain shallow and loosely consolidated. The field exhibits permeability of up to 0.5 Darcies<sup>1</sup> and oil viscosity reaching 600 mPa s. Currently, the oilfield is experiencing a drop in production, with reservoir pressure maintained through water injection. Sand production was observed in all Lower Cretaceous Formations (horizons A, B, D, G, and V) but not in Jurassic Formations (J1 and J2). The oilfield has more than 7000 wells, including injecting wells and abandoned wells. Production and geological data from 2000 producing wells were analyzed in the current study. Figure 2 shows the locations of wells (blue cross) in the field along with fault lines (yellow lines).

<sup>1</sup> 1 darcy =  $9.869233 \times 10^{-13} \text{ m}^2 = 0.9869233 \text{ } \mu\text{m}^2$

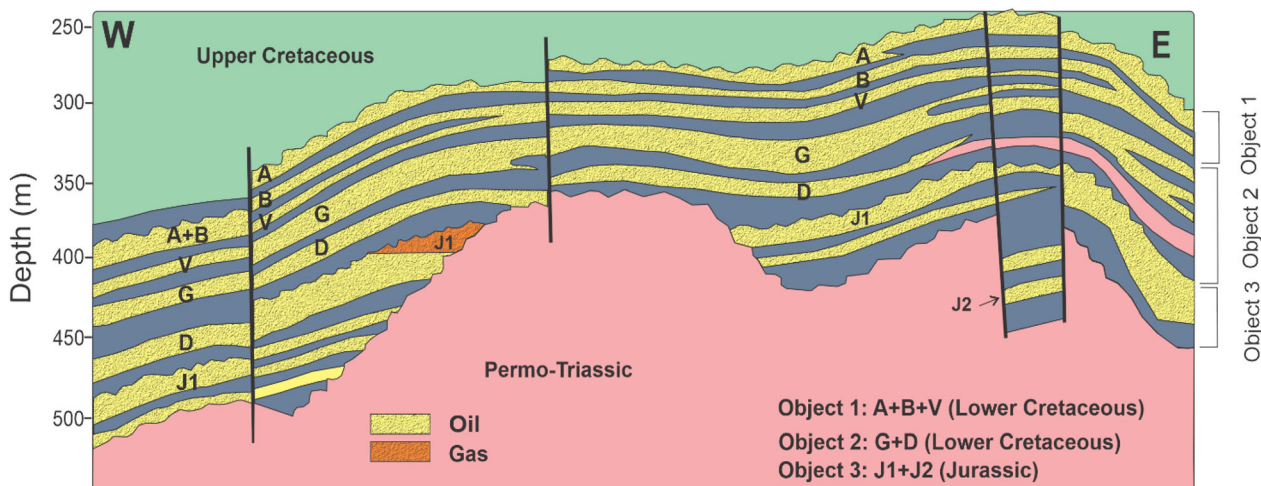


Figure 1. Geological cross-section of the Karazhanbas oilfield (modified after Shabdirova et al. (2023)).

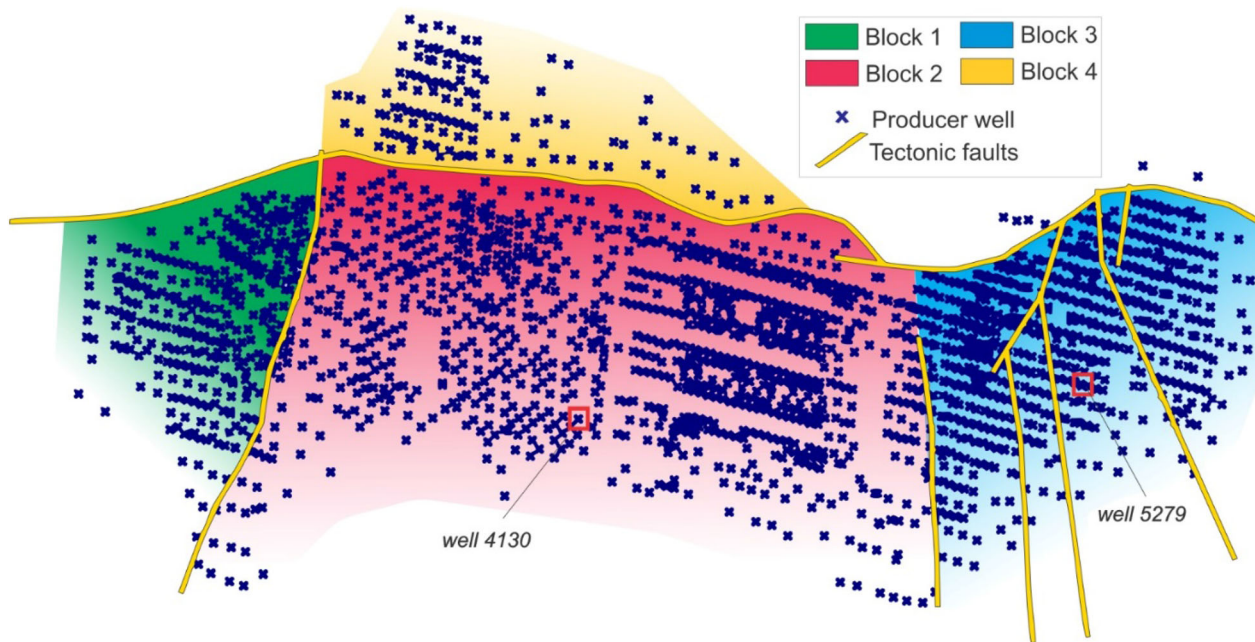


Figure 2. Location of the producing wells in the Karazhanbas oilfield.

The properties of the formations are presented in Appendices 1 and 2. The data suggest significant differences in the composition of these geological layers. Horizon G shows a slightly lower quartz content and a similar range for feldspars and rock fragments compared to other Cretaceous Formations. Kaolinite in the G layer is noted in small amounts and is locally concentrated in the form of

bundles in pore fillings, surrounding grains, and sometimes forming vermiculite booklets. In the A layer and in Jurassic deposits, medium-crystallized kaolinite predominates, covering and replacing grains. Layers A and B are characterized by relatively moderate gas content and saturation pressure, suggesting a less volatile oil type compared to deeper layers. Layers G and D exhibit a broader range

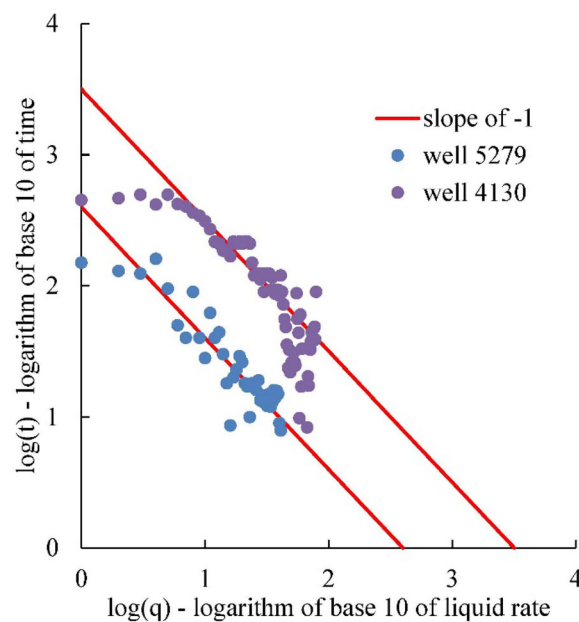
of gas content and saturation pressure, pointing to variability in reservoir conditions or oil types, possibly indicating a more complex geological history or variations in organic matter input and thermal maturation. Lower Cretaceous Formations (A, B, D, G, and V) show high permeability values, particularly horizon G with values up to 4000 mD<sup>2</sup>, suggesting highly conductive pathways for fluid flow, which could contribute to sand production under production stress.

The area of the oilfield is divided into four blocks, separated by tectonic faults. The wells situated closer to a fault are believed to have a boundary-dominated regime (Ortiz, 2013). However, flow regime analysis of the flowing wells in different blocks showed that even the wells located far from the faults also have a boundary-dominated regime (Fig. 3). The flow regime analysis was based on a logarithmic plot of the liquid rate over time of the flowing well, which in this case had a slope of -1, indicating that well production was affected by the boundary (Ortiz, 2013), which is a fault in this case.

## METHODOLOGY

### Data Collection

The production data from an oilfield come as a database containing monthly information for each well, such as production date, producing horizon, liquid rate, sand mass, reservoir pressure, and water cut. These data were refined by removing any missing or irrelevant information. Reservoir thickness and depth were sourced from well perforation intervals, while porosity and net-to-gross ratios were obtained from well logs. However, these logs were available for only about 10% of the wells. For each geological horizon, these values were averaged. For wells lacking specific data, the averaged values from their respective horizons were used as estimates. As a result, for ML modeling, selected inputs included date, liquid rate, water cut, reservoir thickness, depth, porosity, and net-to-gross ratio. The output variable chosen was sand production. A separate file with the wells' coordinates and a map with faults were used to define the shortest distance of each well to a fault.



**Figure 3.** Example of the flow regime of the flowing well, located farthest from the fault: The log-log curve has an approximate slope of -1.

### Machine Learning Algorithms

This research examined one classification task and four regression tasks. For the classification task, ML algorithms such as RF, SVM, decision tree (DT), and K-nearest neighbors (KNN) were utilized. In the case of regression tasks, the RF algorithm was chosen based on its superior performance in previous sand production prediction studies (Shabdirova et al. 2023). While prior research indicates the RF algorithm as the most effective for analyzing sand production, its advantage over alternative models is not markedly significant. More crucially, the quality of input data plays a vital role, underscoring the importance of data processing. Consequently, this investigation shifted the emphasis from algorithm selection to enhancing data analytics practices.

### Model Training and Validation

In this study, we sought to address the following main research questions:

1. Is it possible to predict sand production in wells within a specific block using data from other blocks?

<sup>2</sup> 1 mD = 9.869233 × 10<sup>-16</sup> m<sup>2</sup>

2. Can knowledge of sand production from certain geological horizons be applied to predict sand production in wells from different horizons?
3. What impact does the quality of input data have on prediction accuracy?
4. How does the distance of wells from faults influence sand production predictions?

To answer the first question, multiple ML simulations with various training and testing data combinations were conducted. For example, in one simulation, a model was trained using data from wells in Block 1 and then tested using data from wells in Block 2. Another simulation involved training with Block 1 data and testing with Block 3 data, and so forth.

Addressing the second question involved training and testing models with data segmented by geological horizons within the same block, specifically Block 2, chosen for its high well count. In an initial simulation for this scenario, the model was trained on data from horizon A and tested on data from horizon B within Block 2, followed by subsequent tests on data from other horizons like G and D.

For the third question, three simulations were conducted using the same set of testing data, which comprised data from a selection of wells in Block 2 representing 20% of the total. The training data varied: The first simulation included only operational data (liquid rate and water cut), the second used only geological data (reservoir thickness, depth, porosity, and net-to-gross ratio), and the third combined both operational and geological data.

In response to the fourth question, wells in Block 2 were categorized as 'Near' or 'Distant' based on their proximity to faults—less than 500 m and more than 500 m away, respectively. In the Karazhanbas field, no investigations have been conducted to assess the fault damage zone around the faults, and so the exact distance from the fault that remains unaffected by near-fault damages is unknown. For this study, we used a distance of 500 m based on the study by Choi et al. (2016), who analyzed the existing literature on fault damage zones for different types of reservoirs and found that the damage zone around faults does not exceed 500 m in all reported cases. The ML model training and testing were performed using different combinations of these categories. When models were trained and tested within the same category ('Near'/'Near' and 'Distant'/'Distant'), the data were split

such that 80% were used for training and 20% for testing.

For all tasks, the root mean square error (RMSE) served as the primary validation metric. Additionally, in the fourth task, the  $R^2$  (coefficient of determination) metric was also utilized alongside RMSE for a more comprehensive assessment of model performance.

### Input Variable Analysis

An analysis of the input variables in terms of box-and-whisker plots and correlation plots is shown in Figure 4. The box-and-whisker plots (Fig. 4a) illustrate the diversity in the input variables. The plot's upper and lower boundaries represent the third and first quartiles, respectively. The horizontal line within each plot is the median for each variable. These plots are beneficial for identifying outliers and verifying the data prior to modeling. In the current dataset, a notable number of outlier points were observed for the 'liquid rate' and 'sand' variables.

Correlation between variables can be assessed through the correlation coefficient, which ranges from -1 to +1. The strength of a relationship between two variables is shown by the absolute value of this coefficient. A positive or negative sign denotes a direct or inverse relationship, respectively. In this study, the overall correlation among variables appears to be low. The highest correlation, with coefficient of 0.37, was found between 'ntg' (net-to-gross ratio) and 'porosity.' It is important to highlight, however, that there is no inherent physical relationship between net-to-gross and porosity of a rock, suggesting that the observed correlation might not imply a direct causality. There was a positive correlation of 0.10 between liquid rate and sand production, suggesting that, as the liquid rate increases, sand production also tends to increase, albeit weakly. This finding contrasts with conventional analytical sand prediction models, which typically indicate a direct proportionality between liquid rate and sand production. Another unexpected result from this ML analysis was the negligible correlation of 0.011 between reservoir thickness and sand production. Traditional analytical and numerical models generally predict that larger reservoirs produce more sand. Similarly, while higher porosity is often associated with increased sand production in conventional models, our dataset shows a very weak correlation of 0.041 between porosity and sand

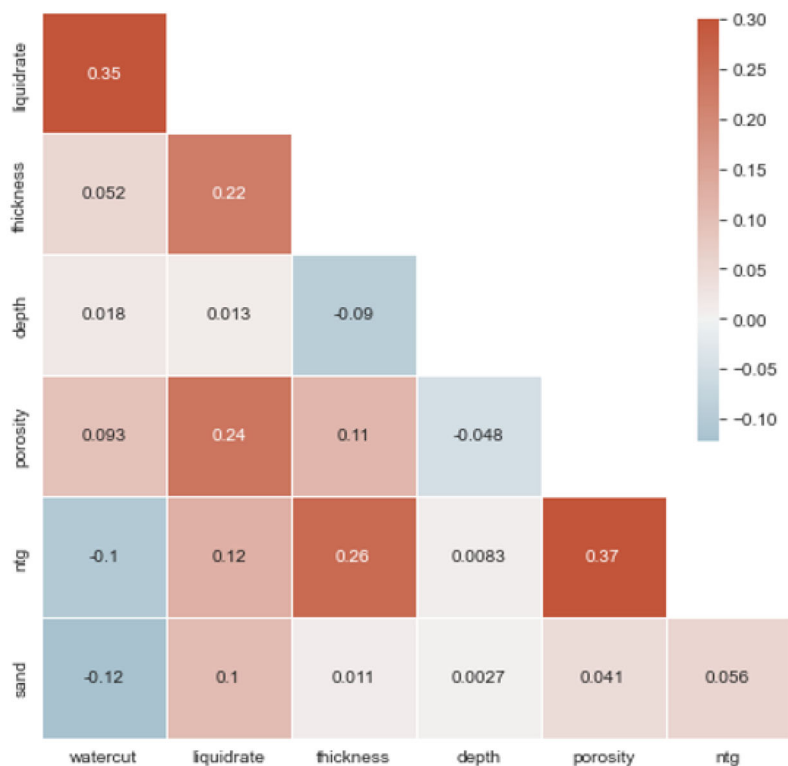
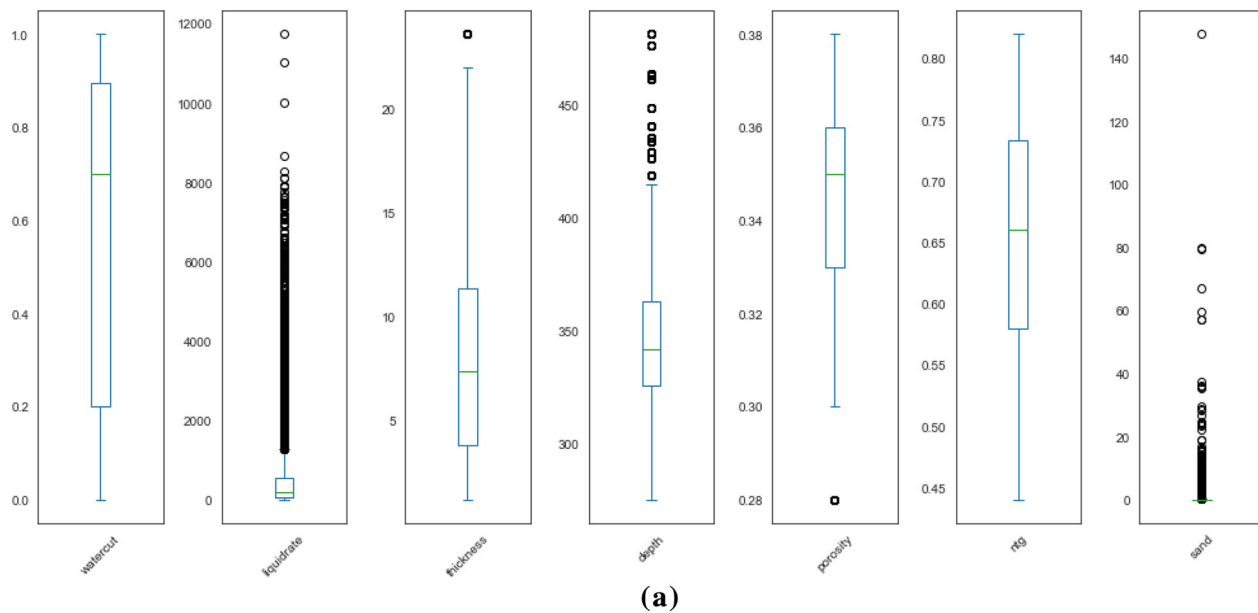


Figure 4. Input variables analysis: (a) box-and-whisker plots and (b) correlation matrix.

production, suggesting minimal influence. Depth is another important parameter for sand production as it represents stress, which is crucial for rock failure. Additionally, the compartmentalized geology of the field results in different depths for the same formation, adding to the complexity. Despite the low linear correlation between depth and sand production, we can expect a more complex relationship that can be revealed through ML modeling.

## RESULTS AND DISCUSSION

The ML analysis started with application of different ML methods to classify the wells into sand-producing wells and no sand-producing wells based on a threshold value of 0.1 t/month, which was established by the operating company. Table 1 shows the accuracy (Forsyth, 2019) of each method, which is good for all ML methods.

Figure 5 shows the locations of the wells with sand production of more than 0.1 t/month, i.e., sand-producing wells. Most of them are located in near-fault regions. The color scale corresponds to the average sand production of each well. While most wells produce sand at less than 5 t/month, several wells have extremely high sand production (red and yellow dots in the plot). The well data suggest that all the wells produce from horizons A and/or G (Fig. 1) and with very high liquid rate up to 2000 t/day. Figure 6 plots the average flow rate values of randomly selected 100 wells. Red and blue bars correspond to sand-producing and no sand-producing wells, respectively. Thus, the results indicate that high sand production was associated with high flow rates.

### Is it Possible to Predict Sand Production in Wells within a Specific Block using Data from Other Blocks?

Further analysis involved assessing the correlation between different blocks (Fig. 2) in terms of sand production prediction. Block 2 had the largest

area and, correspondingly, the largest number of wells and production data. The simulations were conducted to evaluate whether the production data of all sand-producing wells in one block can be used to predict the data from another block. Figure 7 shows the RMSE of the ML modeling for different training and testing data combinations. Generally, models trained on a more diverse set of blocks (e.g., 2 + 3 + 4, 2 + 3) tended to perform better (lower RMSE) than those trained on fewer blocks. This suggests that including a wider range of data in the training set improves prediction accuracy.

The performance of the models varied significantly depending on the test block used. For instance, models tested on Block 1 generally had lower RMSEs compared to those tested on Blocks 3 or 4. This could indicate differences in the characteristics of each block or the quality and quantity of data available from each block. Higher RMSEs were consistently observed for certain test blocks (like Block 4) regardless of the training data combination. This might point to inherent challenges in modeling the data from these blocks, possibly due to complex geological features, variability in data quality, or other factors.

As expected, the best results were achieved for the cases with large amounts of training data. However, further studies demonstrated that the principle of 'more data yielding better outcomes' does not uniformly apply.

### Can Knowledge of Sand Production from Certain Geological Horizons be Applied to Predict Sand Production in Wells from Different Horizons?

The study was then narrowed to Block 2 wells, and further analysis was focused on assessing the correlation between different horizons. The observations from Figure 8, however, present a complex picture; the data do not readily offer clear insights into the correlation between horizons, making it challenging to draw definitive conclusions. Notably, horizon G consistently emerged as the most problematic, frequently registering the highest RMSEs across all training sets, indicating its complexity. On a positive note, the analysis indicated that integrating multiple training horizons generally led to a reduction in RMSE. This trend underscores the potential benefits of diversifying training data to improve model accuracy. Looking ahead, incorporating the mechanical properties of each horizon as

**Table 1.** Classification results

ML method	Classification accuracy
RF	0.81
SVM	0.87
DT	0.81
KNN	0.86

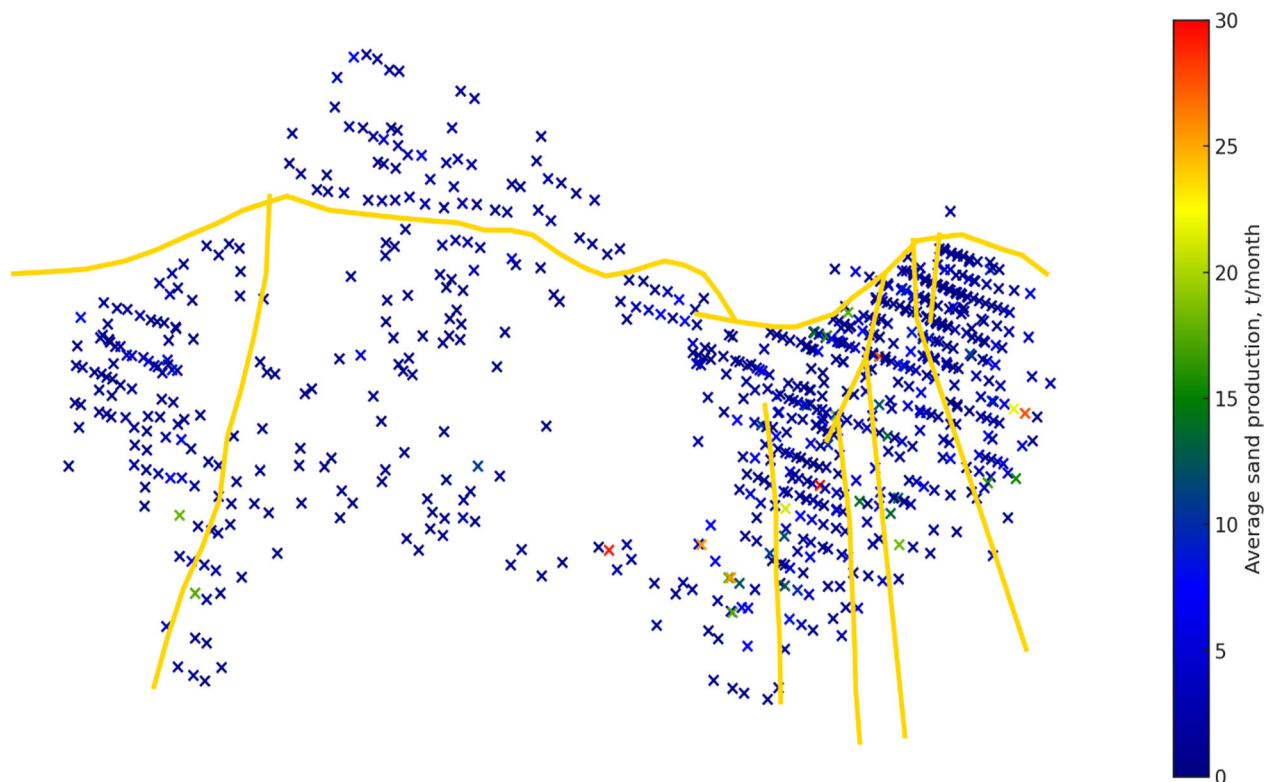


Figure 5. Locations of the sand-producing wells with sand production of more than 0.1 t/month.

input data in future studies could provide a more nuanced understanding and potentially elevate the precision of the predictions.

#### What Impact does the Quality of Input Data Have on Prediction Accuracy?

As presented above, selected inputs included date, liquid rate, water cut, reservoir thickness, depth, porosity, and net-to-gross ratio. Liquid rate and water cut can be considered as operational input data, while reservoir thickness, depth, porosity and net-to-gross ratio represent geological features. ML simulations were conducted to assess impact of input data type on the prediction. Figure 9 presents exemplary outcomes based on a specific set of training data. The RMSE was the lowest (0.85) when the input data contained only operation data: flow rate and water cut. This can be attributed to the predominance of the quality rather than the quantity of input data: Using all available input data results in a higher RMSE of 1.12 compared to just operational

data. This increase implies that including more types of data does not necessarily improve model performance and might introduce noise or irrelevant information. Thus, porosity and net-to-gross ratio values were sourced from logs, available for only one-tenth of the wells. These values were then averaged for each horizon. For the remaining wells without specific data, values were extrapolated from the corresponding horizon. Moreover, geological data remained unchanged over the production. Even in instances where it might vary, data for such changes were not available. These two factors introduce noise into the input data, resulting in diminished accuracy of the prediction results.

#### How does the Distance of Wells from Faults Influence Sand Production Predictions?

Further analysis was focused on assessing the impact of fault proximity on the predictive accuracy of the sand production prediction model. The analysis was categorized into six distinct cases, each

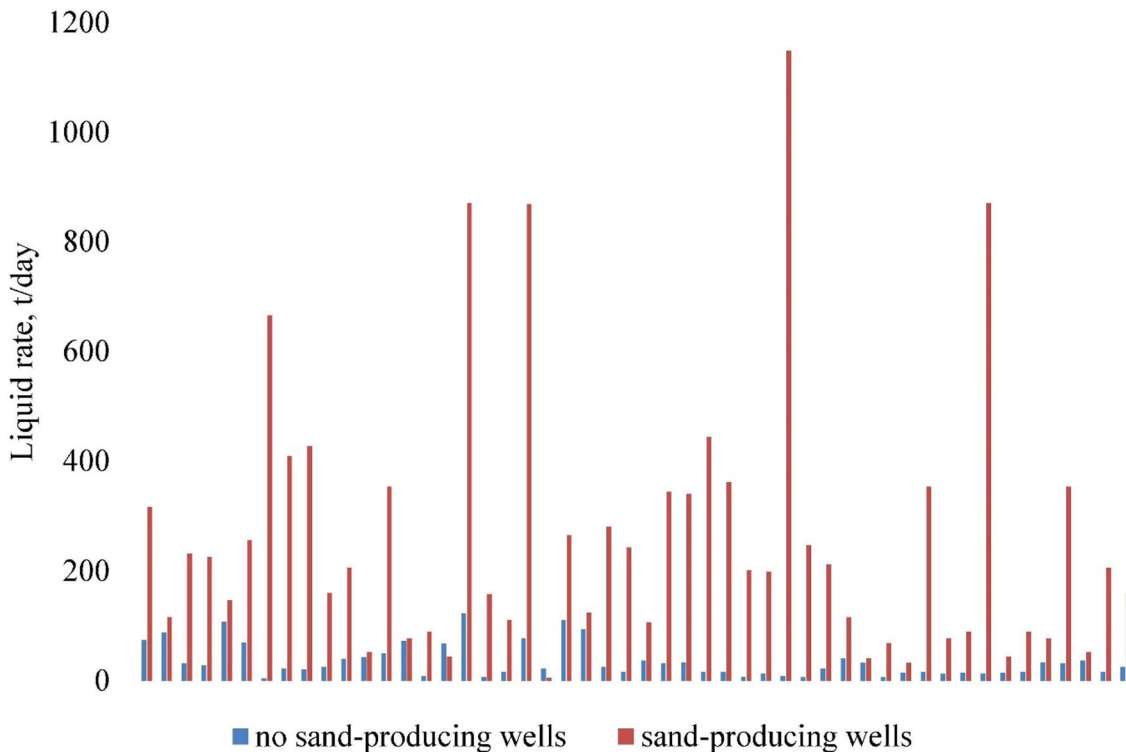


Figure 6. Average liquid rate of the sand-producing and no sand-producing wells.

representing a combination of training and testing scenarios based on the proximity of wells to geological faults. 'Distant' wells were situated more than 500 m away from all faults on the margins of Block 2, while 'Near' wells were located within 500 m of these faults. The simulations demonstrated the complexities inherent in model evaluation when multiple key metrics were used, specifically RMSE and  $R^2$  (Fig. 10). An exclusive reliance on  $R^2$  suggests a moderate level of model performance, with values clustering between 0.4 and 0.6 across different train-test combinations. Conversely, RMSE offers a more discerning perspective, indicating optimal prediction performance when training and testing datasets comprised wells of similar locations (0.05 for distant/distant train/test combination). This divergence underscores the inherent limitations of traditional metrics like RMSE and  $R^2$ , particularly in capturing the critical nuances of sand production, such as the significance of peak values. These peak values are paramount, given their substantial operational implications.

Given the operational importance, not only of the magnitude but also of the timing of peak sand production, it is essential to anticipate these high-risk events for effective operational planning and risk management. Consequently, we advocate for the adoption of new metrics tailored to the unique demands of sand production prediction. These proposed metrics, peak prediction accuracy (PPA) and peak time prediction accuracy (PTPA), were designed to precisely gauge a model's proficiency in forecasting these operationally pivotal events. By focusing on these aspects, PPA and PTPA offer a more operationally relevant assessment of model performance, aligning the evaluation more closely with the end-users' needs and the practical demands of sand production management. Thus:

$$PPA = 100 - \frac{|\text{actual peak sand mass} - \text{predicted peak sand mass}|}{\text{actual peak sand mass}} \times 100$$

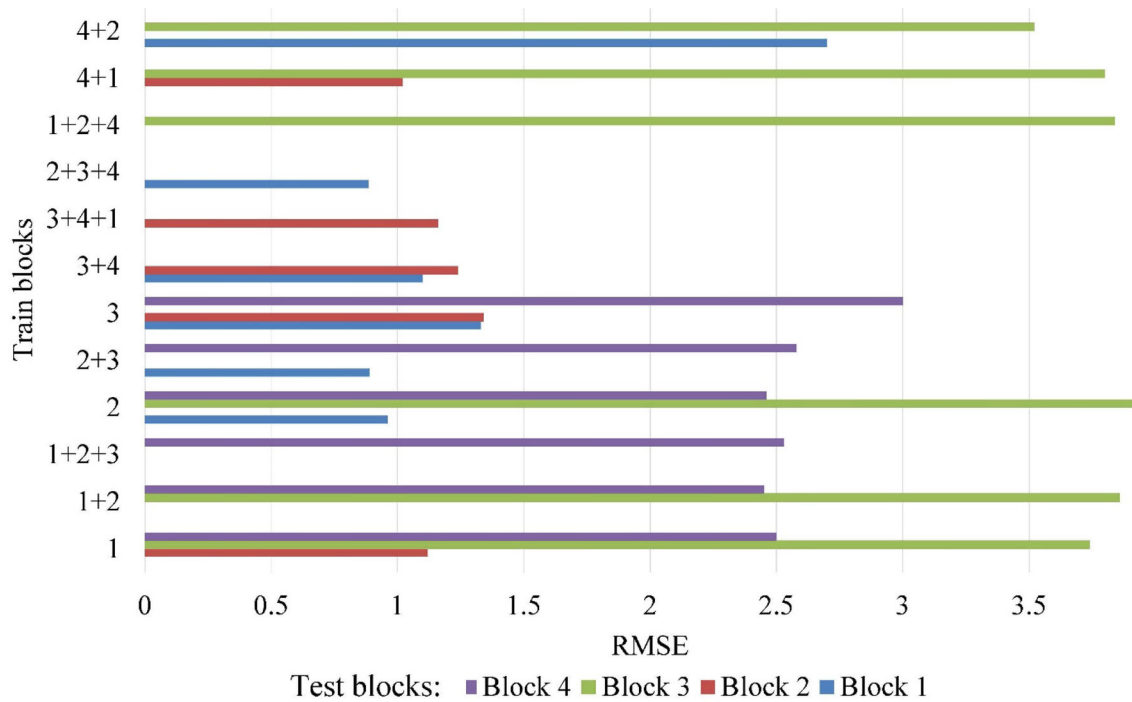


Figure 7. ML prediction results for different train and test blocks.

$$PTPA = 100 - \frac{|\text{time at actual peak sand mass} - \text{time at predicted peak sand mass}|}{\text{time at actual peak sand mass}} \times 100$$

Figure 11 shows bar charts with the new metrics. It becomes evident that predicting the timing of peak sand production events poses significant challenges. The data suggest a more reliable identification of sand production patterns in wells distant from faults. However, wells in proximity to faults seem to be subject to unique geological and operational influences that significantly affect sand production. This underscores the complexity introduced by fault proximity, necessitating tailored approaches in predictive modeling to account for these specific factors.

Incorporating stress-related domain-specific knowledge and relevant features into the model could improve prediction accuracy. In summary, despite the fact that all wells in the field are produced in a boundary-dominated regime, the proximity to faults appears to be a significant factor influencing the prediction of sand production events. Tailoring models to specific conditions, enriching training data with diverse scenarios, and incorpo-

rating domain-specific insights are key steps toward improving the prediction accuracy for both the magnitude and timing of sand production events.

### CONCLUSIONS

This study presented an extensive analysis of sand production prediction in the Karazhanbas oilfield, leveraging ML techniques to improve operational efficiency and risk management. By employing a rich dataset from 2000 wells, the study delved into the intricate dynamics of sand production, highlighting challenges in the prediction of sand production including the peak sand mass and its occurrence. As an oilfield usually covers very large areas with thousands of wells over long production life spanning several decades, it produces a large amount of data. While operational data can be monitored continuously, geological data are much more limited with discrete samples, some parameters of which could only be estimated as average values over certain horizons and locations.

ML usually requires a large dataset for training to improve prediction accuracy. Geological and operational data, however, were not always mea-

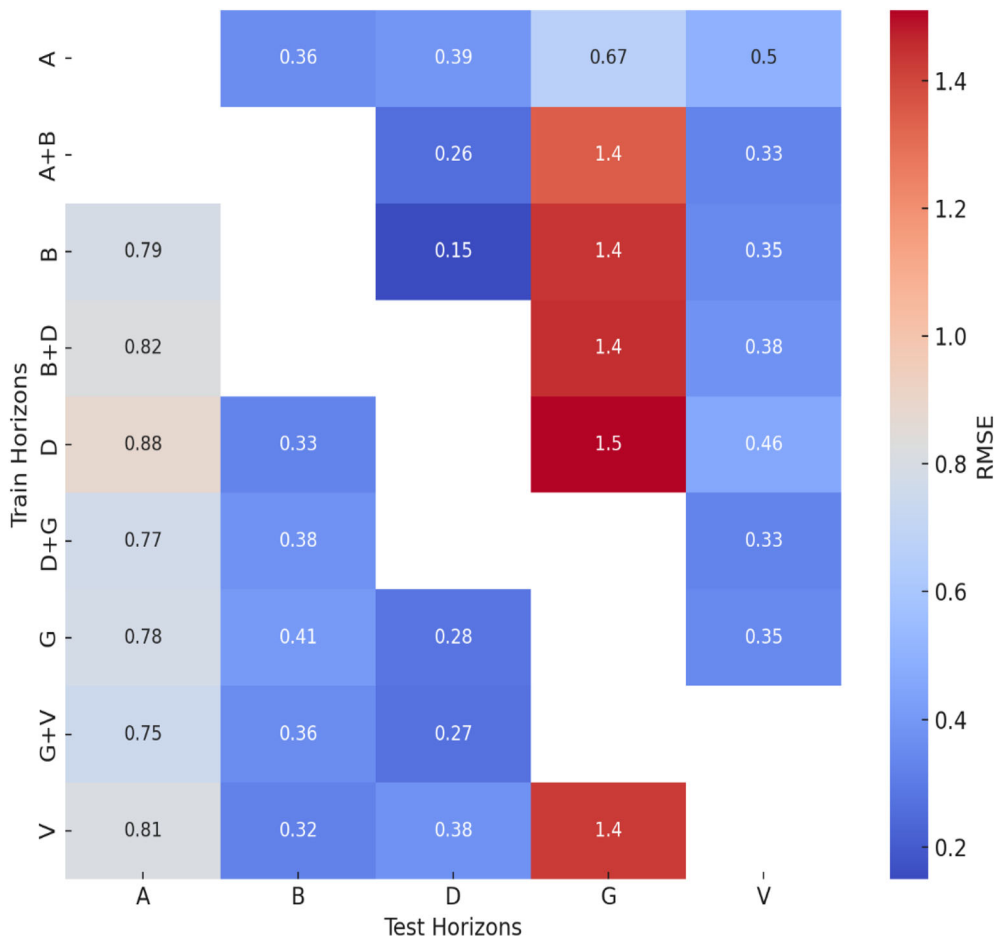


Figure 8. Prediction results for different combinations of training and testing horizons.

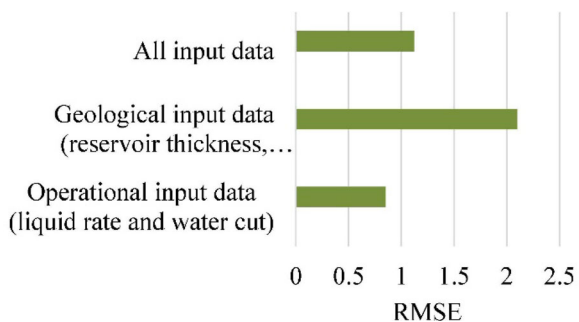


Figure 9. Impact of input data type on the prediction performance.

sured leading to imbalanced dataset. Data processing was found very important for a meaningful prediction of sand production across the oilfield with the ML techniques. Different scenarios were con-

ducted to investigate the sensitivity of the prediction accuracy depending on the location and depth of the data sources as well as on the proximity to the nearby major geological faults. Incorporating mechanical property of each horizon as input data in future studies could provide a more nuanced understanding and potentially elevate the precision of the predictions. This finding elucidates the complex interplay between geological and operational factors in sand production, advocating for models tailored to specific conditions and enriched with diverse data scenarios and domain-specific insights.

The investigation focused on several key research questions:

- The analysis demonstrated that ML models trained on diverse datasets from multiple blocks performed better than those trained on fewer

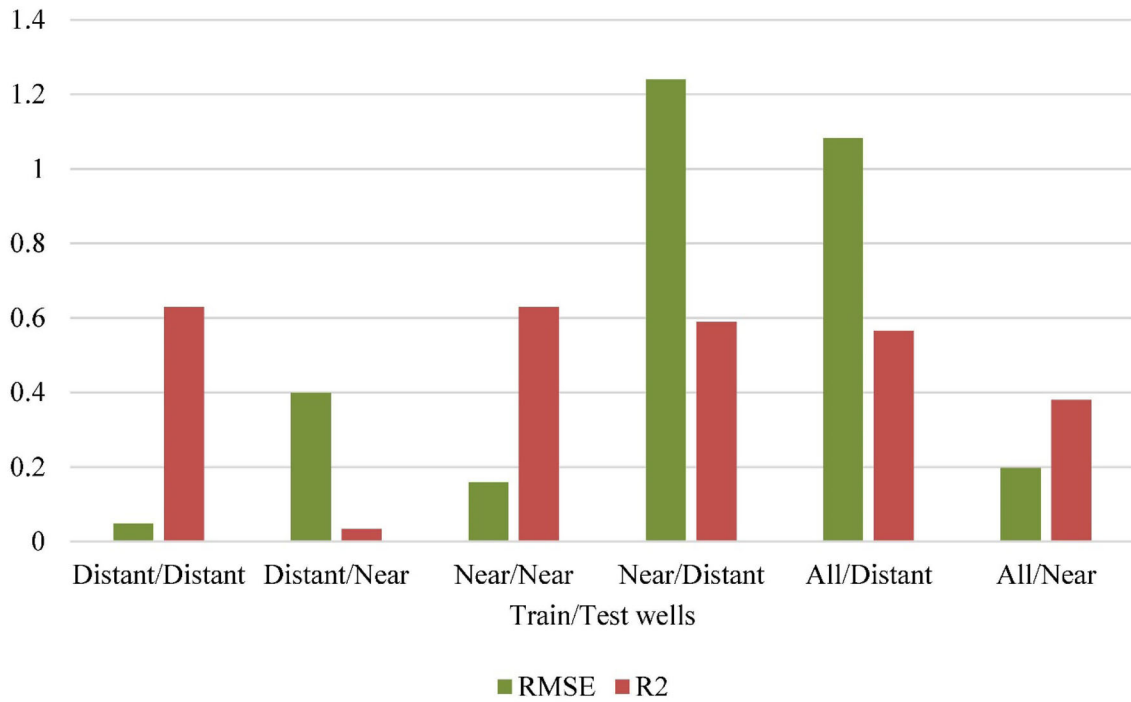


Figure 10. Impact of fault proximity on prediction performance.

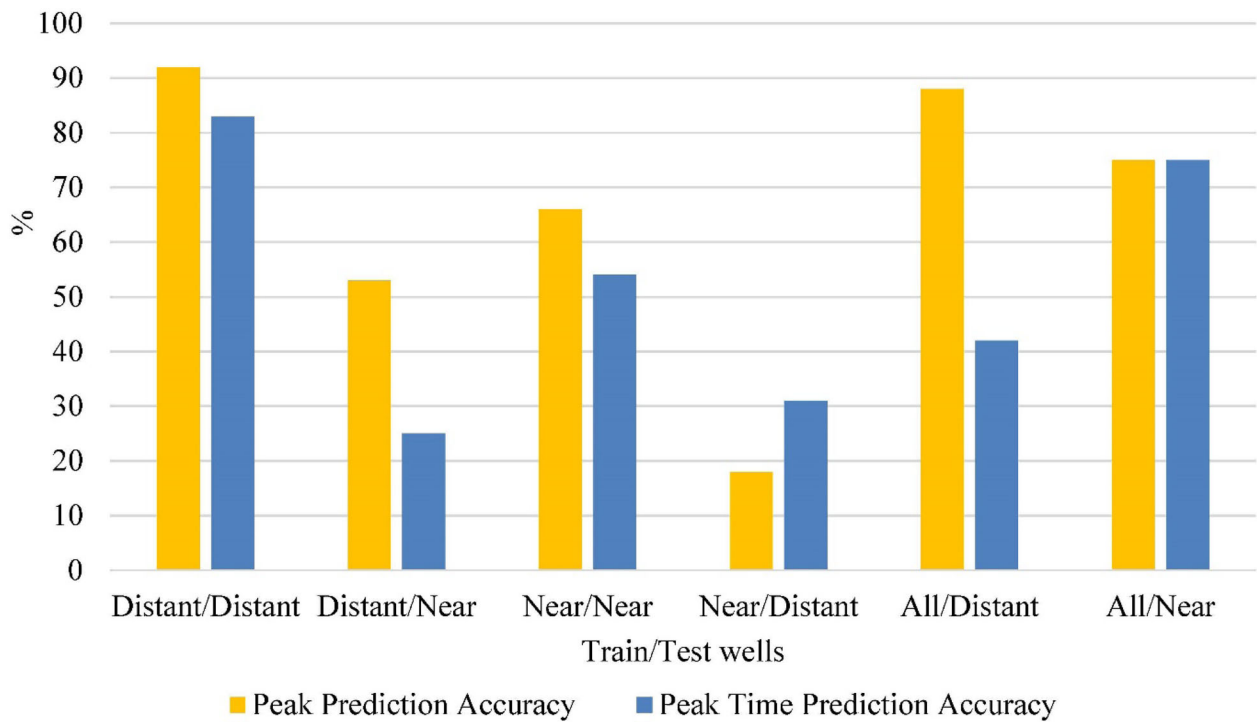


Figure 11. Sand production prediction performance with new metrics. Note: Higher accuracy is associated with values closer to 100%.

blocks, highlighting the importance of data diversity in improving prediction accuracy.

- The study found that integrating data from multiple horizons generally reduced the prediction error, suggesting that a more comprehensive dataset enhances prediction accuracy. However, some horizons consistently showed higher RMSEs, indicating their complexity and the need for additional data or refined models.
- The findings revealed that using only operational data (liquid rate and water cut) resulted in lower RMSE than using all available data, including geological features. This suggests that the quality of input data significantly impacts prediction accuracy, and introducing noise through less accurate geological data can diminish model performance.
- The study underscored that wells' proximity to faults significantly affects sand production predictions. Models trained and tested on wells from similar locations (e.g., distant from faults) performed better, emphasizing the need for tailored approaches in predictive modeling to account for these geological complexities.

The applied ML methods demonstrated commendable classification accuracy in identifying sand-producing wells. The research introduced novel metrics—PPA and PTPA—tailored to the operational needs of sand production management. These metrics were instrumental in evaluating model performance more relevantly, especially in forecasting operationally pivotal events like peak sand production.

In conclusion, this research marks a significant stride in sand production prediction. It not only reinforces the potential of ML in enhancing understanding and forecasting of sand production, but it also paves the way for future studies to focus on refining prediction models. By incorporating stress-related domain-specific knowledge and enriching training data, future research can further improve the accuracy and reliability of sand production prediction, ensuring more informed decision-making

and efficient operational management in oilfield operations.

## ACKNOWLEDGMENT

This research was supported by the following grants: 1) Ministry of Education and Science of the Republic of Kazakhstan grants No. AP13068648. 2) Nazarbayev University CRP grant No. 11022021CRP1506. The APC was paid by Nazarbayev University CRP grant No. 11022021CRP1506.

## DECLARATIONS

**Conflict of Interest** The authors wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

## OPEN ACCESS

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## APPENDIX

See Tables 2 and 3.

Table 2. Average parameters of reservoir oil

Item	Number		Range of change	Average
	Wells	Samples		
Horizon A—north				
Saturation pressure, MPa	3	4	1.10–2.96	1.85
Gas factor, m <sup>3</sup> /t	3	4	6.49–12.00	9.09
Volume factor at Psat, fr of unit	3	4	1.007–1.016	1.013
Viscosity, MPa*s	3	4	304–500	411
Density, g/cm <sup>3</sup>	3	4	0.8904–0.9388	0.9211
Horizons A, B—central, west				
Saturation pressure, MPa	9	11	1.14–2.19	1.72
Gas factor, m <sup>3</sup> /t	9	11	3.60–8.20	5.97
Volume factor at Psat, fr of unit	8	10	1.003–1.028	1.014
Viscosity, MPa*s	8	10	182–576	378
Density, g/cm <sup>3</sup>	8	10	0.9160–0.9389	0.9282
Horizons A, B, V—east				
Saturation pressure, MPa	5	10	1.18–2.09	1.59
Gas factor, m <sup>3</sup> /t	5	10	2.42–7.74	5.29
Volume factor at Psat, fr of unit	5	10	1.007–1.016	1.012
Viscosity, MPa*s	5	10	311–805	541
Density, g/cm <sup>3</sup>	5	10	0.9235–0.9373	0.9292
Horizons G, D—central, west				
Saturation pressure, MPa	14	18	1.20–2.42	2.03
Gas factor, m <sup>3</sup> /t	14	18	3.90–10.09	7.07
Volume factor at Psat, fr of unit	13	17	1.001–1.035	1.020
Viscosity, MPa*s	11	14	171–663	449
Density, g/cm <sup>3</sup>	14	18	0.9140–0.9560	0.9304
Horizons G, D—east				
Saturation pressure, MPa	8	14	1.59–2.27	1.87
Gas factor, m <sup>3</sup> /t	8	14	5.95–12.15	8.21
Volume factor at Psat, fr of unit	8	14	1.014–1.024	1.019
Viscosity, MPa*s	8	14	387–600	448
Density, g/cm <sup>3</sup>	8	14	0.9184–0.9301	0.9234
Horizon J1—west				
Saturation pressure, MPa	1	1	–	1.12
Gas factor, m <sup>3</sup> /t	1	1	–	4.70
Volume factor at Psat, fr of unit	1	1	–	1.011
Viscosity, MPa*s	1	1	–	500
Density, g/cm <sup>3</sup>	1	1	–	0.9329
Horizons J1, J2—central				
Saturation pressure, MPa	1	2	–	2.16
Gas factor, m <sup>3</sup> /t	1	2	–	7.15
Volume factor at Psat, fr of unit	1	2	–	1.021
Viscosity, MPa*s	1	2	–	530
Density, g/cm <sup>3</sup>	1	2	–	0.9228
Horizons J1, J2—east				
Saturation pressure, MPa	5	9	1.50–2.27	1.94
Gas factor, m <sup>3</sup> /t	5	9	6.79–12.15	9.07
Volume factor at Psat, fr of unit	5	9	1.014–1.024	1.019
Viscosity, MPa*s	5	9	387–600	458
Density, g/cm <sup>3</sup>	5	9	0.9184–0.9277	0.9232

**Table 3.** Porosity and permeability from core measurements

Horizon	Number of wells	Samples	Porosity				Permeability $\times 10^{-3}$ . mcm <sup>2</sup>			
			Average	Min	Max	Coeff. of variation	Average	Min	Max	Coeff. of variation
A	19	94	0.324	0.262	0.423	0.0124	980.9	24.50	10000.0	3.85
B	12	39	0.322	0.262	0.389	0.0157	294.3	27.58	2460.0	2.24
V	19	51	0.310	0.262	0.400	0.0108	244.0	22.70	2540.0	2.33
G	52	346	0.319	0.261	0.418	0.0104	761.7	21.20	8377.0	2.71
D	12	94	0.311	0.261	0.436	0.0106	723.4	21.14	3394.2	1.34
J1	27	392	0.325	0.260	0.413	0.0115	1465.9	20.60	10000.0	2.56
J2	11	218	0.320	0.260	0.393	0.0082	607.6	20.92	4697.0	1.26

## REFERENCES

- Abdelghany, W. K., Hamed, 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>.
- Al-Shaabi, 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.
- Chen, F. (2012). Coupled flow discrete element method application in granular porous media using open source codes. *Uma Ética Para Quantos?*, XXXIII(2), 81–87.
- Choi, J., Edwards, P., Ko, K., & Kim, Y. S. (2016). Definition and classification of fault damage zones, a review and a new methodological approach. *Earth-Science Reviews*, 152, 70–87.
- Cui, Y., Nouri, A., Chan, D., & Rahmati, E. (2016). A New Approach to DEM Simulation of Sand Production. *Journal of Petroleum Science and Engineering*, 147, 56–67.
- 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.
- Fjaer, E., Holt, R. M., Horsrud, P., Raaen, A. M., & Risnes, R. (2008). *Petroleum Related Rock Mechanics* (Vol. 30). Elsevier B.V.
- Forsyth, D. (2019). *Applied Machine Learning*. Springer.
- 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>.
- Garolera, D., Carol, I., & Papanastasiou, P. (2019). Micromechanical analysis of sand production. *International Journal for Numerical and Analytical Methods in Geomechanics*, 43(6), 1207–1229.
- Geilikman, M. B., Dusseault, M. B., & Dullien, F. A. (1994). Sand production as a viscoplastic granular flow. *SPE Symposium on Formation Damage Control*. <https://doi.org/10.2118/27343-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>.
- Gholami, R., Aadnoy, B., Rasouli, V., & Fakhari, N. (2016). An analytical model to predict the volume of sand during drilling and production. *Journal of Rock Mechanics and Geotechnical Engineering*, 8(4), 521–532.
- 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.
- Han, Y., & Cundall, P. (2017). Verification of two-dimensional LBM-DEM coupling approach and its application in modeling episodic sand production in borehole. *Petroleum*, 3(2), 179–189.
- Hayavi, M. T., & Abdideh, M. (2017). Establishment of tensile failure induced sanding onset prediction model for cased-perforated gas wells. *Journal of Rock Mechanics and Geotechnical Engineering*, 9(2), 260–266.
- 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(04), 261–273.
- 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*. <https://doi.org/10.2118/26541-ms>.
- Ketmalee, T., & Bandyopadhyay, P. (2018). Application of neural network in formation failure model to predict sand production. In *Offshore Technology Conference Asia 2018, OTCA 2018*, pp. 1–10. <https://doi.org/10.4043/28506-ms>.
- Khamehchi, E., RahimzadehKivi, I., & Akbari, M. (2014). A novel approach to sand production prediction using artificial intelligence. *Journal of Petroleum Science and Engineering*, 123, 147–154.
- Khamitov, F., Minh, N. H., & Zhao, Y. (2022). Numerical investigation of sand production mechanisms in weak sandstone formations with various reservoir fluids. *International Journal of Rock Mechanics and Mining Sciences*, 154, 105096.
- Kim, Sh., Sharma, M., & Fitzpatrick, H. (2011). A predictive model for sand production in poorly consolidated sands. *International Petroleum Technology Conference*. <https://doi.org/10.2523/15087-MS>.
- 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*, 13(1), 154–166.
- Li, X., Feng, Y., & Gray, K. E. (2018). A hydro-mechanical sand erosion model for sand production simulation. *Journal of Petroleum Science and Engineering*, 166, 208–224.
- Morita, N., Whitfill, D. L., Fedde, O. P., & Lovik, T. H. (1989). Parametric study of sand-production prediction: Analytical approach. *SPE Production Engineering*, 4(01), 25–33. <http://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 *Society of Petroleum Engineers - SPE Nigeria Annual International Conference and Exhibition 2021, NAIC 2021*, 1–16. <https://doi.org/10.2118/207193-MS>.

- Nouri, A., Vaziri, H., Belhaj, H., & Islam, R. (2006a). 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.
- Nouri, A., Vaziri, H., Belhaj, H. A., & Islam, M. R. (2006b). Sand-production prediction, a new set of criteria for modeling based on large-scale transient experiments and numerical investigation. *SPE Journal*, 11(02), 227–237.
- Ortiz, L. (2013). *Pressure normalization of production rates improves forecasting results*. Master Thesis. Texas A&M University, 2013. <https://doi.org/10.2118/168974-ms>.
- Papamichos, E., & Furui, K. (2013). Sand Production Initiation Criteria and Their Validation. In *47th US Rock Mechanics/ Geomechanics Symposium*.
- 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., Minh, N. H., & 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*. <https://doi.org/10.1007/s11053-023-10234-z>.
- Shabdirova, A., Minh, N. H., & Zhao, Y. (2019). A sand production prediction model for weak sandstone reservoir in Kazakhstan. *Journal of Rock Mechanics and Geotechnical Engineering*. <https://doi.org/10.1016/j.jrmge.2018.12.015>.
- 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, 565–75.
- 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. *Energies*, 15. <https://doi.org/10.3390/en15186509>.
- Wang, H., Gala, D. P., & Sharma, M. M. (2019a). Effect of fluid type and multiphase flow on sand production in oil and gas wells. *SPE Journal*, 24(2), 733–743.
- Wang, H., Yang, X., Zhang, W., & Sharma, M. M. (2018). Predicting sand production in HPHT wells in the Tarim Basin. In *Proceedings - SPE Annual Technical Conference and Exhibition 2018-Septe* (December 2019), 24–26. <https://doi.org/10.2118/191406-ms>.
- Wang, M., Feng, Y. T., Zhao, T. T., & Wang, Y. (2019b). Modelling of sand production using a mesoscopic bonded particle lattice Boltzmann method. *Engineering Computations (Swansea, Wales)*, 36(2), 691–706.
- 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/10.2118/96-70>.
- 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(07), 596–600.
- Wu, B., Choi, S. K., Denke, R., Barton, T., Viswanathan, C., Lim, S., Zambari, M., & Shaffee, S. (2016). A new and practical model for amount and rate of sand production. *Offshore Technology Conference*, 18.
- Yi, X. (2003). *Numerical and Analytical Modeling of Sanding Onset Prediction*. Texas A&M University.