Geologic Sweet Spot Identifier: A Machine Learning Approach to Optimal Drilling Locations

Aditya Holla, Ameera Aslam, Beshoy Shaker, Elijah Flores, Jayant Bhaskarun

CONTENTS

[1.0 ABSTRACT 4](#_Toc211977326)

[2.0 INTRODUCTION 4](#_Toc211977327)

[2.1 Context 4](#_Toc211977328)

[2.2 Problem 5](#_Toc211977329)

[2.3 Response 6](#_Toc211977330)

[3.0 BACKGROUND 7](#_Toc211977331)

[3.1 Prior Research in Sweet Spot Identification 7](#_Toc211977332)

[3.2 Machine Learning Applications in Reservoir Characterization 8](#_Toc211977333)

[4.0 METHODS 9](#_Toc211977334)

[4.1 Munging 9](#_Toc211977335)

[4.1.1 Missing and Zero Values 9](#_Toc211977336)

[4.1.2 Outlier Detection and Transformation 9](#_Toc211977337)

[4.1.3 Feature Deletion 9](#_Toc211977338)

[4.2 Feature Engineering 10](#_Toc211977339)

[4.2.1 Variable Renaming and Encoding 10](#_Toc211977340)

[4.2.2 Correlation and Multicollinearity Handling 10](#_Toc211977341)

[4.2.3 Normalization Decisions 11](#_Toc211977342)

[4.3 Analysis 11](#_Toc211977343)

[4.3.1 Model Development and Validation 11](#_Toc211977344)

[4.3.2 Sweet Spot Score Formulation 12](#_Toc211977345)

[4.3.3 Spatial Visualization 13](#_Toc211977346)

[5.0 RESULTS 13](#_Toc211977347)

[5.1 Sweet Spot Score Distribution 13](#_Toc211977348)

[5.2 3D Interpolated Map of Reservoir Productivity 14](#_Toc211977349)

[6.0 DISCUSSION/INTERPRETATION 15](#_Toc211977350)

[6.1 Geological Implication of High-Score Zones 15](#_Toc211977351)

[6.2 Reservoir Connectivity and Anisotropy 15](#_Toc211977352)

[6.3 Compartmentalization and Flow Barriers 15](#_Toc211977353)

[6.4 Model Limitations and Sources of Uncertainty 16](#_Toc211977354)

[7.0 CHALLENGES AND RESOLUTIONS 17](#_Toc211977355)

[7.1 Data Limitations 17](#_Toc211977356)

[7.2 Feature Redundancy and Correlation 17](#_Toc211977357)

[7.3 Outlier and Scaling Issues 18](#_Toc211977358)

[7.4 Overfitting Mitigation 18](#_Toc211977359)

[7.5 Reproducibility and Future Plans 18](#_Toc211977360)

[8.0 CONCLUSION 19](#_Toc211977361)

[9.0 REFERENCES 20](#_Toc211977362)

[10.0 FIGURE CAPTIONS 20](#_Toc211977363)

# 1.0 ABSTRACT

Oil and gas exploration remains one of the most data-intensive yet uncertain stages of the production process. Determining optimal drilling locations, or “sweet spots,” traditionally relies on geological interpretation and seismic analysis, methods that can be subjective and limited in scalability. This study applies machine learning to identify sweet spots by analyzing a well-level dataset provided by ConocoPhillips, which includes 55 wells and 14 variables describing petrophysical, production, and spatial characteristics such as porosity, permeability, and depth. Data preprocessing involved median imputation for missing values, outlier transformation, and one-hot encoding of categorical features. Regression models supported by cross-validation were used to explore relationships among key variables and to guide the development of a composite Sweetspot Score. This score integrates geological and production parameters into a single, interpretable metric that quantifies reservoir productivity potential. The results were visualized through a three-dimensional Sweetspot Map that revealed zones of enhanced reservoir quality, highlighting areas with strong permeability, porosity, and production performance. The findings emphasize the potential of data-driven workflows to improve drilling accuracy, reduce costs, and enhance reservoir characterization. Future work will focus on incorporating geospatial learning and uncertainty quantification to strengthen predictive robustness and field applicability.

# 2.0 INTRODUCTION

## 2.1 Context

Oil and gas exploration plays a vital role in meeting global energy demands, yet it remains one of the most uncertain and expensive stages in the production process. Determining where to drill has traditionally relied on geological surveys, seismic imaging, and expert interpretation. While these methods provide valuable insights, they often involve large amounts of complex data that are difficult to analyze objectively, which can lead to subjective decisions.

In recent years, advances in data science and machine learning have created new opportunities to analyze subsurface data more systematically. By identifying patterns within geological, petrophysical, and production data, machine learning models can support more accurate and efficient exploration strategies that help companies target the most promising drilling zones while reducing cost and risk.

Previous research and industry practice have shown that geological and petrophysical properties such as porosity, permeability, depth, and facies are strong indicators of oil productivity. Models using these features can reasonably estimate how much oil a well might produce, especially when combined with historical production data. Studies have also demonstrated that machine learning techniques, such as regression and ensemble methods, can improve prediction accuracy compared to traditional statistical approaches. However, most of these models focus on predicting well performance rather than identifying where, geographically, the most productive areas lie (Ibrahim et al., 2022; Du et al., 2024).

This study builds on existing research in petroleum engineering and machine learning applied to reservoir characterization. While previous studies have explored the use of statistical and machine learning methods for predicting productivity, porosity, and permeability individually, few have integrated these features to identify sweet spots directly. Using well-level datasets such as ConocoPhillips’ Sweetspot Dataset, this study extends previous methods by incorporating more advanced machine learning approaches to detect areas of high potential. This aligns with current trends in geoscience and energy research where predictive modeling is increasingly used to reduce exploration costs, optimize well placement, and improve decision-making.

## 2.2 Problem

Determining the optimal drilling locations, or “sweet spots,” is a crucial step in petroleum reservoir development. Traditional methods for selecting drilling sites rely heavily on expert interpretation of geological and seismic data, which can be subjective, time-intensive, and difficult to scale. With the growing availability of well-level geological and production data, there is an opportunity to leverage machine learning to uncover complex, nonlinear relationships between subsurface features and production outcomes.

This research aims to develop a machine learning-based sweet spot identification model using a dataset provided by ConocoPhillips, which includes well-level data such as bottomhole coordinates, porosity (POROS), permeability (KX, KY), total depth (TD), fluid and gas production metrics (Co [MSTB], Cw [bbl], Cg [mmcf]), and geological facies classifications. The model will use these variables to predict regions with the highest likelihood of oil productivity, as indicated by cumulative production (Co [MSTB]).

Although machine learning offers a promising way to identify sweet spots, existing models often treat each well as an independent data point. This limits their ability to account for how geological characteristics vary spatially across a field. Without incorporating spatial relationships, models cannot identify patterns or clusters that indicate the most productive zones. In addition, the small dataset used in this study, containing only 55 wells, restricts the model’s ability to generalize to new drilling locations. Sparse data and uneven spatial coverage introduce uncertainty and can lead to overfitting.

A key knowledge gap therefore remains in how to integrate geospatial relationships directly into predictive modeling. While variables such as porosity and permeability are well understood indicators of productivity, the spatial interactions between wells and their geological context remain difficult to capture. Without incorporating this information, drilling decisions may continue to rely on incomplete data or expert judgment, leading to higher costs and inefficiencies in exploration.

## 2.3 Response

Despite the dataset’s limited size, this research serves as an initial step toward developing data-driven methods that can guide drilling decisions and improve exploration efficiency. Regression was chosen as the initial modeling approach because it produces interpretable results and is well-suited for smaller datasets. This method allows the model to identify which geological and petrophysical factors most strongly influence production, providing a foundation for data-informed decision-making in drilling operations.

In future work, the study aims to expand beyond simple regression by incorporating geospatial learning. The goal is to develop a deep learning model capable of analyzing spatial relationships between wells and predicting high-potential drilling zones across an entire area rather than only for individual wells. By combining production data with spatial variables such as bottomhole coordinates and geological facies, this advanced model would be able to identify and visualize sweet spots that represent the most promising locations for new drilling.

Additionally, spatial mapping can be used to compare oil wells across different variables, providing insight into the physical factors that influence productivity. This helps explain why certain patterns appear in the data and supports the development of models tailored specifically to the characteristics of the studied reservoir rather than generalized cases.

This two-step approach, beginning with regression and progressing toward spatial deep learning, allows the study to build a strong analytical foundation while acknowledging current data limitations.

# 3.0 BACKGROUND

## 3.1 Prior Research in Sweet Spot Identification

Previous research in petroleum engineering has focused extensively on understanding the geological and petrophysical factors that control hydrocarbon productivity. Studies have shown that rock properties such as porosity, permeability, facies, and total depth are reliable indicators of reservoir quality and fluid flow potential. Early approaches relied primarily on statistical regression and petrophysical interpretation to identify high-productivity zones, but these methods were limited by the complexity and heterogeneity of subsurface data.

Ibrahim et al. (2022) demonstrated that predictive modeling can capture production behavior across diverse formations using long-term regression experiments on oil, gas, and water outputs. Similarly, Du et al. (2024) applied deep learning to predict oil production in U.S. reservoirs, achieving improved accuracy compared to traditional empirical models. These studies provided valuable insight into the relationship between geological properties and production performance but were primarily concerned with predicting output at existing wells rather than identifying new drilling opportunities.

More recent research has sought to combine petrophysical and spatial parameters to locate areas of enhanced reservoir potential, often referred to as “sweet spots.” However, most models still evaluate wells individually and do not integrate coordinate-based spatial data that could reveal geographic clusters of productivity. This lack of spatial awareness limits the applicability of existing methods to field-scale exploration planning. The current study builds upon this prior work by combining geological, production, and spatial features within a unified modeling framework designed to map zones of high production potential.

## 3.2 Machine Learning Applications in Reservoir Characterization

Machine learning has become an increasingly valuable tool for reservoir characterization because of its ability to identify nonlinear relationships and process large, multidimensional datasets. Techniques such as regression, random forests, and neural networks have been used to estimate reservoir properties, predict production trends, and automate data interpretation. These approaches have demonstrated higher predictive accuracy than conventional statistical methods, especially when working with incomplete or noisy data.

Lavi et al. (2024) introduced data-driven imputation strategies to handle missing production records, showing that machine learning can effectively recover incomplete measurements and maintain data integrity. Li et al. (2024) further demonstrated that integrating multiple features such as permeability, porosity, and depth within a weighted productivity model improves reservoir evaluation accuracy. Their hierarchy of control factors (K > φ > So > PV > Pr > TLV > FD > PR) established permeability and porosity as the dominant drivers of productivity, providing a foundation for multi-variable scoring frameworks such as the one developed in this study.

Machine learning applications in reservoir engineering continue to evolve toward combining predictive modeling with spatial analysis. By incorporating well coordinates and depth-dependent features, models can move beyond isolated well prediction to spatial mapping of reservoir potential. This study advances that direction by applying regression-based learning to quantify the relationships among petrophysical and production variables and then visualizing the results through a three-dimensional Sweetspot Map.

This study builds upon established research in petroleum engineering and data-driven modeling while addressing gaps identified in prior work. To provide context for the methods used, the following section reviews existing studies on sweet spot identification and highlights recent advancements in machine learning applications for reservoir characterization. This background establishes the foundation for the analytical framework developed in this project.

# 4.0 METHODS

This section outlines the workflow used to process, analyze, and model the dataset provided by ConocoPhillips. The process includes three major stages: data munging, feature engineering, and analysis. Each stage addresses specific challenges such as missing values, outlier handling, multicollinearity, and model interpretability.

## 4.1 Munging

### 4.1.1 Missing and Zero Values

The data munging process begins with reading and inspecting the dataset to identify missing values, outliers, and redundant information. Out of the fifty-five wells in the dataset, five (Wells 8, 9, 27, 28, and 47) contain zero values in the oil and gas production columns. Before altering these values, a literature review was conducted to ensure the approach was appropriate.

According to Lavi et al. (2024), missing or zero data in petroleum field datasets may result from a well shutdown during data collection, equipment malfunction, or data corruption during transfer. Missing data should therefore be treated as valid operational results rather than random omissions. To avoid information loss and reduce bias, **median imputation** was applied because it is more robust to outliers than mean imputation. This method preserves the small dataset and stabilizes the production variable distributions.

### 4.1.2 Outlier Detection and Transformation

Outlier detection was performed using the **interquartile range (IQR)** method. As shown in Figure 3, water production exhibited strong positive skewness with several extreme values. To reduce variance and minimize the influence of these high-magnitude points, a data transformation was applied.

Because most water production values were near zero, a standard logarithmic transformation would have produced negative results. To avoid this, a **log(x + 1)** transformation was used instead. This approach effectively normalized the distribution, reducing the impact of outliers while maintaining the geological relevance of the variable.

In addition, redundant identifiers such as **well name** and **well number** were removed since each dataset row already provides unique well identification. The **future pressure prediction** column was also excluded to prevent data leakage, as it would act as a predictive variable in a predictive problem.

### 4.1.3 Feature Deletion

Columns containing redundant or non-predictive information were removed. The **well name** and **well number** fields were dropped since the row index already provided unique well identification. To avoid data leakage, the **future pressure prediction** variable was excluded because including it in model training would bias the results and reduce generalizability.

## 4.2 Feature Engineering

### 4.2.1 Variable Renaming and Encoding

After munging, feature engineering prepared the dataset for modeling. All variable names were standardized using **snake case** for clarity and consistency (for example, “Bottomhole X” was renamed bh\_x, and “Cumulative Oil” to oil\_prod\_mstb). The categorical variable **FACIES** was one-hot encoded into binary columns, allowing it to be used in regression analysis.

### 4.2.2 Correlation and Multicollinearity Handling

Statistical testing identified variables with limited predictive power, such as facies\_5, total\_depth\_md, and water\_prod\_bbl, which were removed due to low significance. Correlation analysis revealed strong relationships between **permeability** and **porosity**, and between **depth** and **past pressure**, indicating potential multicollinearity. These findings informed the decision to use **regularized regression models**, such as ridge or lasso regression, that can penalize redundant predictors while retaining essential geological information (Figure 1).

### 4.2.3 Normalization Decisions

Normalization was selectively applied to maintain proportional feature contributions during modeling. Production variables such as oil and gas outputs were normalized using **min–max scaling**, while categorical and dimensionless variables were kept in their encoded format to preserve interpretability.

Prior to integration, all variables were normalized to a dimensionless scale between 0 and 1 using min–max normalization:

This ensures that all parameters contribute proportionally regardless of their native units or magnitude differences.

## 4.3 Analysis

### 4.3.1 Model Development and Validation

The initial modeling phase used **Ordinary Least Squares (OLS) regression** to evaluate relationships among key reservoir parameters, including permeability (perm\_x, perm\_y), porosity (φ), oil production (oil\_prod\_mstb), and gas production (gas\_prod\_mmcf). The OLS model served as a diagnostic tool for assessing linear relationships between petrophysical properties and production outcomes. This method was selected for its transparency, computational simplicity, and interpretability, providing a useful baseline for identifying influential variables.

However, the OLS results demonstrated limited predictive robustness due to nonlinear interactions within the dataset. Residual analysis revealed heteroscedasticity and spatial autocorrelation, violating the assumption of independent and identically distributed residuals. Additionally, permeability anisotropy between perm\_x and perm\_y introduced nonlinear flow effects that could not be captured by a linear model.

A further limitation involved the **oil production variable (oil\_prod\_mstb)**, which displayed strong correlations with multiple predictors. Although oil production is a critical measure of reservoir performance, using it directly as a predictor risked overfitting, as it already reflects the combined influence of permeability, porosity, and completion efficiency. This issue highlighted the need for a new modeling framework that could maintain interpretability while reducing sensitivity to production-driven noise.

### 4.3.2 Sweet Spot Score Formulation

To address these limitations, a **composite Sweetspot Score** was developed. This score preserves the physical interpretability of regression outputs while reducing statistical overfitting. It integrates permeability, porosity, and production indicators into a single metric representing overall reservoir potential.

Weights were assigned to each contributing variable based on both literature findings and empirical relationships observed in the data. Following Li et al. (2024), permeability and porosity were emphasized as the dominant drivers of productivity, with lesser weights given to production and depth-related terms.

The Sweetspot Score was computed using a weighted linear combination of normalized parameters:

Sweetspot Score = 0.35K + 0.25ϕ + 0.20So ​+ 0.15Pr ​+ 0.05(1−D)

where **K** represents mean permeability, **φ** represents porosity, **Sₒ** represents oil production, **Pᵣ** represents gas production, and **D** represents normalized total depth. The inverse scaling of depth accounts for productivity loss with increasing depth due to compaction and reduced flow capacity.

This composite scoring method reduces dependence on a single variable, balances geological and production influences, and provides a physically meaningful way to identify areas of high productivity potential.

### 4.3.3 Spatial Visualization

The computed Sweetspot Scores were visualized spatially using bottomhole coordinates (bh\_x, bh\_y) and total depth to represent three-dimensional well positioning. A **3D scatter plot** was generated, where color intensity indicated the Sweetspot Score magnitude.

This visualization revealed clear clustering of high-scoring wells, forming distinct “sweet spot” zones within the reservoir. These spatial patterns aligned with known geological trends, indicating that areas of high permeability and porosity also corresponded to higher production potential.

The spatial mapping step bridged the analytical and geological perspectives of the study. By connecting data-driven insights with physical locations, it allowed for a more practical interpretation of model outputs and served as a foundation for future spatial modeling approaches.

# 5.0 RESULTS

## 5.1 Sweet Spot Score Distribution

A three-dimensional interpolation of the **Sweetspot Score** was created to visualize the spatial variation of reservoir productivity across the study area (Figure 2). The Sweetspot Score, calculated as a weighted linear combination of permeability, porosity, production, and depth parameters, provides a unified measure of reservoir quality and flow potential. The adopted weighting scheme (0.35 for mean permeability, 0.25 for porosity, 0.20 for oil production, 0.15 for gas production, and 0.05 for total depth) was adapted from the productivity control hierarchy proposed by Li et al. (2024), which identifies permeability and porosity as dominant controls on well output.

The resulting Sweetspot Score distribution differentiates high- and low-productivity wells across the dataset. Higher scores generally occur within intermediate depth intervals and regions exhibiting stronger permeability and porosity values. In contrast, lower scores are concentrated in deeper or low-permeability zones where compaction and diagenesis likely reduce pore connectivity.

Overall, the Sweetspot Score successfully integrates petrophysical and production variables into a single interpretable metric that reflects the underlying physical properties governing hydrocarbon flow.

## 5.2 3D Interpolated Map of Reservoir Productivity

The three-dimensional **Connected Sweetspot Map** (Figure 2) combines borehole spatial coordinates (bh\_x, bh\_y) with total measured depth to visualize the Sweetspot Score as a color-mapped surface. Linear interpolation converts discrete well data into a continuous surface, revealing both lateral and vertical continuity in reservoir productivity potential.

Warmer colors (yellow to red) represent higher Sweetspot Scores and indicate zones where permeability and porosity combine with favorable production metrics, suggesting enhanced hydrocarbon flow potential. Cooler colors (blue to purple) correspond to lower-quality zones with reduced permeability or porosity.

Several high-score clusters are visible, laterally connected at intermediate depths. These areas likely correspond to depositional or structural features such as channel sands or high-permeability stratigraphic units that promote greater fluid connectivity. Shallower and mid-depth regions consistently exhibit stronger scores than deeper sections, where compactional effects and diagenetic cementation may diminish permeability.

The resulting visualization highlights spatially coherent regions of high reservoir potential and provides a foundation for interpreting geological trends that control productivity.

# 6.0 DISCUSSION/INTERPRETATION

## 6.1 Geological Implication of High-Score Zones

The spatial clustering of high Sweetspot Scores aligns with known geological expectations of enhanced porosity and permeability in less compacted strata. These high-score zones likely represent well-sorted, coarser-grained facies associated with higher reservoir quality and improved flow capacity. Their lateral continuity suggests stratigraphic units or depositional trends, such as channelized sands, that sustain production potential across multiple wells.

This geological coherence confirms that the Sweetspot Score captures meaningful subsurface variation rather than random statistical noise. The framework therefore provides a realistic representation of productivity trends consistent with field-scale geological interpretations.

## 6.2 Reservoir Connectivity and Anisotropy

The Sweetspot Map also illustrates anisotropic flow behavior driven by directional differences between permeability in the x and y directions (perm\_x, perm\_y). Such anisotropy reflects depositional fabrics, lamination, or cross-bedding that preferentially enhance fluid flow along one horizontal axis. These directional properties create elongated high-score regions, consistent with permeability anisotropy observed in heterogeneous sandstone reservoirs (Meyer and Krause, 2006; Li et al., 2024).

Anisotropy acts as a key factor influencing reservoir connectivity, with higher Sweetspot Scores marking corridors of enhanced horizontal flow that could represent preferred pathways for hydrocarbon migration.

## 6.3 Compartmentalization and Flow Barriers

Sharp transitions between high- and low-score zones within the 3D visualization likely represent **reservoir compartmentalization**, where lateral pressure communication is restricted. Such compartments often arise from facies changes, stratigraphic pinch-outs, or low-permeability shale barriers (ScienceDirect Topics, 2025).

In this study, high-score regions can be interpreted as well-connected flow units, while adjacent low-score areas likely function as flow baffles or sealed compartments. This interpretation supports the use of the Sweetspot framework as a diagnostic tool for identifying both connected zones and potential barriers to flow within a field.

## 6.4 Model Limitations and Sources of Uncertainty

While the Sweetspot Score framework produces geologically consistent results, several limitations must be recognized. The dataset includes only fifty-five wells, which constrains the model’s ability to fully capture spatial variability. A small sample size increases sensitivity to outliers and may exaggerate trends driven by a few data points.

Additionally, the feature weighting scheme, though grounded in literature, remains somewhat subjective. Future iterations should apply **sensitivity analysis** to test the robustness of weighting factors and validate them against independent datasets. The limited resolution of spatial interpolation also introduces uncertainty, particularly in areas with sparse data coverage where the model relies on extrapolation.

Certain variables, such as **water productivity**, displayed irregular behavior and were sometimes inversely correlated with expected production trends. A more detailed exploratory data analysis (EDA), including feature-specific spatial mapping, is recommended to understand these anomalies and improve model stability.

Despite these constraints, the overall workflow demonstrates that even small datasets can yield physically interpretable insights when combined with transparent modeling and sound geological reasoning.

# 7.0 CHALLENGES AND RESOLUTIONS

## 7.1 Data Limitations

The most significant challenge faced in this project was the limited size of the dataset, which included only fifty-five wells. This small sample restricted the model’s ability to generalize and limited the diversity of geological and petrophysical conditions represented. Sparse data also made the model more sensitive to outliers and potential noise, which could distort results.

To address this issue, the team applied median imputation to retain as many observations as possible without introducing bias from missing or zero values. Cross-validation techniques were used to test the model’s consistency across multiple subsets of the data. These steps ensured that results were not overly dependent on individual wells and that statistical relationships remained stable across the dataset.

## 7.2 Feature Redundancy and Correlation

Another major challenge involved feature redundancy, particularly the strong correlations between permeability and porosity, and between depth and pressure. These correlations can lead to multicollinearity, which inflates variance in regression coefficients and reduces model interpretability.

To mitigate this, correlation matrices were generated to identify highly correlated variables. Non-essential or statistically insignificant features, such as facies\_5, total\_depth\_md, and water\_prod\_bbl, were excluded from the model. In addition, regularized regression techniques such as ridge and lasso regression were considered to penalize redundant predictors and maintain balanced model complexity.

## 7.3 Outlier and Scaling Issues

Outliers presented a significant obstacle, especially within the water production variable, which showed strong positive skewness. These extreme values risked distorting regression relationships and masking meaningful trends in other parameters.

To correct for this, outliers were handled using the interquartile range (IQR) method, and the affected variables were log-transformed using **log(x + 1)** to reduce skewness while retaining physical meaning. Furthermore, normalization techniques were applied to scale production and geological variables to a comparable range, ensuring that no single parameter dominated the model due to unit differences or magnitude disparities.

## 7.4 Overfitting Mitigation

Overfitting was a concern, particularly when the model incorporated production-based features such as oil productivity, which inherently reflect multiple underlying factors. To prevent the model from capturing noise instead of genuine geological relationships, the Ordinary Least Squares (OLS) results were used primarily for diagnostic purposes rather than direct prediction.

The development of the **Sweetspot Score** served as a solution to this challenge. By weighting geological and production factors according to established literature rather than relying solely on model-driven coefficients, the Sweetspot framework reduced overfitting risk while preserving interpretability. The use of cross-validation further ensured that results remained consistent across subsets of the dataset.

## 7.5 Reproducibility and Future Plans

Ensuring reproducibility was another key challenge. The workflow involved multiple stages of data transformation, feature engineering, and model computation, which required consistent documentation to avoid discrepancies between runs.

To maintain reproducibility, all scripts were version-controlled, and intermediate datasets were saved after each transformation step. A GitHub repository was created to track the pipeline from raw data to final visualization, allowing transparent replication of results.

Future plans include expanding this framework to incorporate additional data sources and developing a **geospatial deep learning model** capable of analyzing spatial dependencies directly. This approach would move beyond interpolation to capture three-dimensional geological continuity and improve the accuracy of sweet spot prediction.

# 8.0 CONCLUSION

This study demonstrates the potential of machine learning techniques in identifying high-productivity zones within oil and gas reservoirs. By integrating geological, petrophysical, and production data into a composite **Sweetspot Score**, the model provides a quantitative and interpretable measure of reservoir potential. The resulting three-dimensional visualization effectively highlights spatial trends, showing that areas with high porosity and permeability correspond to enhanced production potential.

Despite the limitations imposed by the small dataset and correlations among key variables, the workflow achieved a balance between interpretability and predictive strength. The incorporation of data cleaning, feature engineering, and regularization improved both the accuracy and transparency of the analysis.

The Sweetspot framework serves as a foundation for future expansion into geospatially aware and economically optimized models. With additional data and advanced spatial algorithms, this approach can evolve into a robust decision-support tool that improves drilling efficiency, reduces exploration risk, and deepens understanding of subsurface reservoir behavior.

# 9.0 REFERENCES

1. **Meyer, R., and Krause, F. F. 2006.** Permeability Anisotropy and Heterogeneity of a Sandstone Reservoir Analogue: An Estuarine to Shoreface Depositional System in the Virgelle Member, Milk River Formation, Writing-on-Stone Provincial Park, Southern Alberta. *Bulletin of Canadian Petroleum Geology* 54 (4): 301–318. <https://doi.org/10.2113/gscpgbull.54.4.301>
2. **ScienceDirect Topics. 2025.** Reservoir Compartmentalization—An Overview. *Earth and Planetary Sciences*. <https://www.sciencedirect.com/topics/earth-and-planetary-sciences/reservoir-compartmentalization>
3. **Li, Q., Guo, K., Gao, X., Zhang, S., Jin, Y., and Liu, J. 2024.** “Productivity Prediction Model of Tight Oil Reservoir Based on Particle Swarm Optimization–Back Propagation Neural Network.” *Processes* 12(9): 1890. <https://doi.org/10.3390/pr12091890>
4. **Ibrahim, N. M., Alharbi, A. A., Alzahrani, T. A., Abdulkarim, A. M., Alessa, I. A., Hameed, A. M., Albabtain, A. S., Alqahtani, D. A., Alsawwaf, M. K., & Almuqhim, A. A. 2022.** Well-performance classification and prediction: Deep learning and machine-learning long-term regression experiments on oil, gas, and water production. *Sensors* 22(14): 5326. <https://doi.org/10.3390/s22145326>
5. **Du, J., Zheng, J., Liang, Y., & Shahzad, K. 2024.** A deep learning-based approach for predicting oil production: A case study in the United States. *Energy* 288: 129688. <https://doi.org/10.1016/j.energy.2023.129688>
6. **Lavi, B., Bertini Jr., J.R., Pires, L.C.O., and Schiozer, D.J. 2024.** “Handling Missing Data on Short-term Oil Rate Forecasting in a Pre-salt Offshore Field Benchmark.” ROG.e 2024: Technical Paper, 1–11. <https://doi.org/10.48072/2525-7579.roge.2024.3726>

# 10.0 FIGURE CAPTIONS

Figure 1: P-values of Variables for Gas Production (mmcf)

**A table of numbers and symbols

Description automatically generated**

Figure 2: Three-dimensional Interpolated Map of the Sweetspot Score

**A graph of a sweetpott

Description automatically generated**

Figure 3: Box Plot of the Cumulative Water Production in Barrels

**A diagram of a box plot

Description automatically generated**