IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 16 October 2023, accepted 22 November 2023, date of publication 23 November 2023,
date of current version 15 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3336623

RESEARCH ARTICLE

Reservoir Property Prediction in the North Sea Using Machine Learning

ABDULRAHMAN AL-FAKIH, (Member, IEEE), SANLINN I. KAKA, AND ARDIANSYAH I. KOESHIDAYATULLAH, (Member, IEEE)

College of Petroleum Engineering and Geosciences, King Fahd University of Petroleum Minerals, Dhahran 31261, Saudi Arabia

Corresponding authors: Abdulrahman Al-Fakih (g202103050@kfupm.edu.sa) and Ardiansyah I. Koeshidayatullah (a.koeshidayatullah@kfupm.edu.sa)

The work of Ardiansyah I. Koeshidayatullah was supported in part by the Saudi Data and Artificial Intelligence Authority (SDAIA)—King Fahd University of Petroleum and Minerals (KFUPM) Joint Research Center in AI under Grant JRCAI-RG-05.

ABSTRACT The North Sea sedimentary basin is characterized by geological complexity, encompassing a wide range of rock types and structures, including multiple reservoirs (carbonates and siliciclastic) with variations in reservoir quality and heterogeneity. These phenomena pose significant challenges for accurately predicting reservoir properties using traditional well log analysis. Moreover, these challenges are further compounded by complex geological conditions and scarcity of available data. Hence, the aim of this study was to address these challenges by applying advanced machine learning techniques within this basin. This study delves into both supervised and unsupervised machine learning approaches to forecast the essential petrophysical properties that are crucial for assessing reservoir quality. These properties encompass total porosity, effective porosity, and shale volume, all derived from well log data originating from the North Sea sedimentary basin. The models were trained using data from four wells consisting of 32,215 data points (80% for training, 10% for testing, and 10% for validation). Furthermore, our study introduced pioneering data-driven preprocessing workflow, encompassing exploratory data analysis, missing data imputation, and outlier detection to improve the performance of the machine learning models. ANN and RF models achieved the best results among the algorithms evaluated, with an average MAE of 0.01, RMSE of 0.01, and R-squared of 0.95 for total porosity, effective porosity, and volume of shale, respectively. These metrics demonstrate that the model can accurately predict reservoir properties in a challenging sedimentary basin, even with limited data availability, enabling reservoir characteristics and field development optimization, particularly in areas where core data are scarce.

INDEX TERMS Reservoir property prediction, machine learning, artificial neural networks, K-nearest neighbors, random forests, decision trees, North Sea sedimentary basin.

NOMENCLATURE

ANN	Artificial Neural Networks.
AutoML	Automated Machine Learning.
DT	Sonic Transit Time.
DTs	Decision Trees.
EDA	Exploratory Data Analysis.
GR	Gamma Ray.
IF	Isolation Forest.
IQR	Interquartile Range.
KNN	K-nearest neighbors.
LOF	Local Outlier Factor.
MAE	Mean Absolute Error.
ML	Machine Learning.
MLT	Machine Learning Techniques.
NPHI	Neutron Porosity.
RF	Random Forest.
RHOB	Bulk Density.
RMSE	Root Mean Square Error.
R^2	Coefficient of Determination.
SVM	Support Vector Machine.

The associate editor coordinating the review of this manuscript and approving it for publication was Wentao Fan.

I. INTRODUCTION

A comprehensive understanding of reservoir properties is crucial for the successful extraction of hydrocarbons from

both mature and frontier hydrocarbon fields. Moreover, precise and reliable forecasts of these properties are imperative to maximize oil and gas output from reservoirs. The accuracy of these predictions plays a crucial role in ensuring that production processes are optimized and efficient [1]. Traditional approaches to reservoir property prediction, utilizing well log data and geological models, can be time-consuming, expensive, costly, and potentially inaccurate in complex geological environments. Machine learning (ML) has emerged as a powerful tool in this domain. ML algorithms, trained on data from well logs, core analysis, and seismic surveys, can discern relationships between reservoir properties and various factors, such as rock type, depositional environment, and fluid composition. Once trained, ML algorithms can be used to predict reservoir properties in new areas with limited data [2].

This study focused on the predictive modeling of reservoir properties in the North Sea sedimentary basin using ML techniques. We collected well log data from six wells, trained the models using data from four wells, and evaluated the models using one well for testing and one well each for validation. The artificial neural network (ANN), K-nearest neighbor (KNN), random forest (RF), and decision tree (DT) algorithms were employed to predict the total porosity, effective porosity, and volume of shale. The performance of the models was assessed using evaluation metrics such as the mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R-squared).

The main objectives of this study are as follows:

- Introduce a novel workflow for predicting reservoir properties in the North Sea sedimentary basin using ML techniques.
- Evaluate the performance of four supervised and unsupervised ML algorithms (ANN, KNN, RF, and DT) for reservoir property prediction. This demonstrates the potential of ML techniques to improve reservoir characterization and optimize the development of oil and gas fields in the Dutch sector of the North Sea.

II. LITERATURE REVIEW

The North Sea region, particularly the Netherlands, possesses extensive petroleum reserves, including proven oil reserves of 140.9 million barrels and natural gas reserves of 1.1 trillion cubic meters [3]. Accurate prediction of reservoir properties is crucial because it can reduce costs, enhance efficiency, and boost the profitability of oil and gas production [4]. Well logging offers valuable insights into subsurface parameters such as porosity, shale volume, and permeability [5]. These parameters are essential for estimating reservoir quality, predicting fluid flow behavior, and optimizing production strategies [6]. Previous studies have used empirical models and conventional statistical methods, including regression analysis, ANNs, and fuzzy logic systems, for predicting porosity and permeability from well logs [7], [8], [9].

Numerous studies have utilized machine learning techniques (MLT) to predict porosity, permeability, and resistivity from well logs, often achieving high levels of accuracy and surpassing traditional models in some cases [10], [11], [12], [13], [14], [15], [16]. However, these studies have primarily been concentrated in fields in outside the Dutch sector of the North Sea. This section reviews the background of reservoir property prediction, with a focus on the recent emergence of MLT and its potential application in the Dutch oil and gas industry. Table 1 demonstrates that ML algorithms can accurately predict reservoir parameters, with certain algorithms exhibiting superior performances in specific regions and reservoir types.

Recent advancements in ML for reservoir property prediction have highlighted diverse methodologies. For instance, a study conducted in a Malaysian brownfield utilized Random Forest (RF) algorithms with well logs and core analysis, achieving an R2 of 85% for porosity and 80% for permeability [17]. This approach, which mirrors our data usage, differs in its application to ML. Another study developed a systematic ML approach for reservoir identification and production prediction by employing seven ML methods, including XGBoost, and achieved up to 99% accuracy for effective reservoirs [18]. This contrasts with our study's exclusive focus on property predictions. Additionally, research on reservoir prediction under coastal conditions has employed binary classification algorithms and data augmentation techniques, providing a unique perspective compared to our North Sea focus [19]. These studies collectively underscore the evolving role of ML in reservoir property prediction and contextualize our work.

Discussing the specific applications of ML in various regional contexts [17], [18], [19] leads to the acknowledgment of recent advancements in the field. The implementation of automated machine learning (AutoML) for more efficient and accurate reservoir characterization represents a significant shift in this landscape. Such advancements, as exemplified by the hierarchical AutoML approach used in Alberta's Athabasca Oil Sands [20], offer promising future research directions even though they have not been extensively applied in the Dutch North Sea context.

III. DATA COLLECTION AND PREPROCESSING

A. EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) techniques, including the use of kernel density estimation (KDE) plots and histograms, were employed to gain insights into the dataset characteristics and detect any patterns or anomalies. KDE plots provide a visualization of the estimated probability density function of the data, whereas histograms represent the frequency distribution of the data. Before conducting predictive modeling, EDA was performed to gain insights into the dataset. This step is crucial for understanding the characteristics of the data and identifying patterns or anomalies.

Summary statistics for each numeric column in the data frame, including the number of data points, mean, standard

TABLE 1. This table compares the performance of different machine learning algorithms for predictingreservoir properties.

Author (s)	Year	Study area	ML algorithms used	Performance metrics	Main Differences	Ref.
Al-Fakih et al.	2023	North Sea	ANN, KNN, RF, DT	MAE=0.01, RMSE=0.01, R^2=0.95	Advanced methods, new area.	
Mubarak et al.	2023	Alberta	AutoML	MAE=1.1-8.09%, RMSE=0.45-1.4%, R^2=76.2-77.76%	AutoML introduction	[20]
Tian et al.	2022	Carbonate	Deep learning	MAE=0.03, RMSE=0.04, R^2=0.85	New methods, reservoir type.	[21]
Otchere et al.	2021	Niger Delta	ANN, SVM	MAE, RMSE, R^2=	Different areas, metrics.	[22]
Ali et al.	2023	UK and Norwegian North Sea	LR, SVM, RF, Trees	MAE= 0.0028, RMSE=0.118, R^2=0.75	Similar area, lower results.	[23]

deviation, minimum and maximum values, and percentile values, were computed for the relevant well-log parameters. Table 2 presents summary statistics for the relevant columns.

This analysis provided a comprehensive overview of the distribution and range of values for each well log parameter, enabling a better understanding of the characteristics of the dataset.

Fig. 1(a) presents the KDE plot of the Gamma Ray (GR) log distribution for the six wells, providing insights into the estimated probability density function of the GR values. This plot revealed a multimodal distribution, indicating the presence of different rock types or lithologies within the well log data.

Fig. 1(b) illustrates the KDE plot of the Bulk Density (RHOB) log distribution, offering a visualization of the estimated probability density function for the RHOB values. Similar to the GR log, the RHOB log distribution also exhibited a multimodal pattern, suggesting the presence of multiple rock types or lithologies within the well log data.

Fig. 1(c) displays a histogram plot of the GR log distribution, representing the frequency distribution of the GR values. This plot further explores the distribution patterns of GR log data, highlighting any notable features or variations.

Fig. 1(d) illustrates a histogram of the RHOB log distribution, capturing the frequency distribution of the RHOB values. This plot provides insight into the distribution characteristics of the RHOB log data, confirming the multimodal distribution observed in the KDE plot.

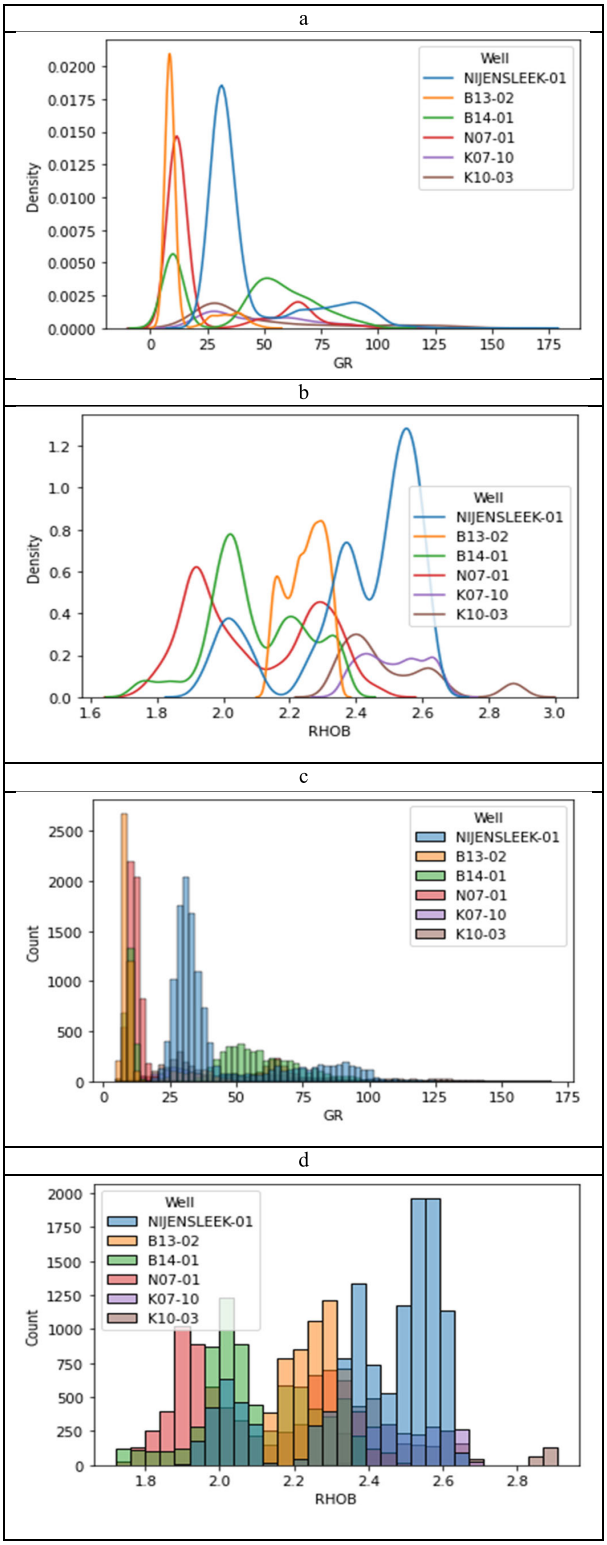

FIGURE 1. Log distribution analysis for the six target wells: (a) KDE plot of GR log distribution and (b) KDE plot of RHOB log distribution. (c) Histogram plot of the GR log distribution and (d) the RHOB log distribution plot.

These visualizations provide initial insights into the distribution and characteristics of well log data. Further analysis and modeling techniques will be employed

TABLE 2. Summary statistics of relevant well log parameters.

PARAMETER	GR (API)	DT (FT, MIN)	RHOB (GM/CM3)	NPHI (PU)
COUNT	36900	36900	36900	36900
MEAN	35.4404	103.3060	2.2727	0.2851
STD	25.7009	35.0914	0.2288	0.1273
MIN	4.5156	45.9502	1.7197	0.0000
25%	11.6291	77.7909	2.0664	0.1748
50%	30.6024	88.9696	2.2939	0.2758
75%	50.4717	124.6676	2.4585	0.3893
MAX	168.8458	200.8674	2.9093	0.5879

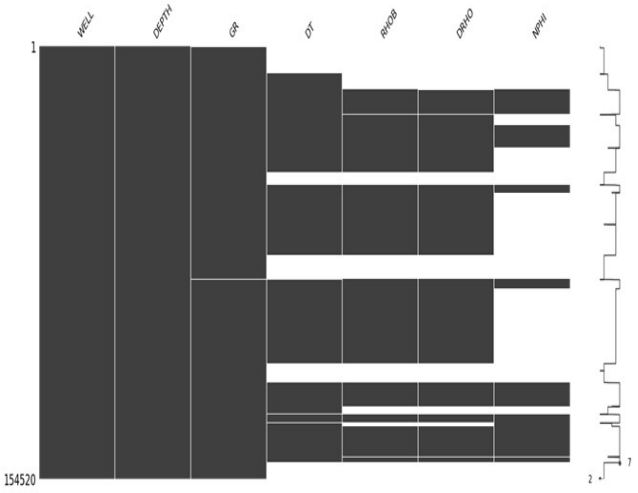

FIGURE 2. Sparkline indicates data completeness for well log features. A complete line at the maximum value denotes data rows with no missing values, ensuring data integrity for analysis and modeling across all six wells.

to explore the relationships between these variables and the target reservoir properties, such as the total porosity, effective porosity, and volume of the shale. By investigating these relationships, we can better understand geological formations and enhance reservoir characterization efforts.

B. DATA IMPUTATION

Addressing missing data, variable discarding and list-wise deletion techniques were employed. Variable discarding removes features with missing values, whereas list-wise deletion eliminates rows with missing data, such as a single depth level for well logs. Excluding the density correction (DRHO) feature improved the data's shape and distribution. To assess the quality of the dataset, Figure 2 shows spark lines indicating data completeness.

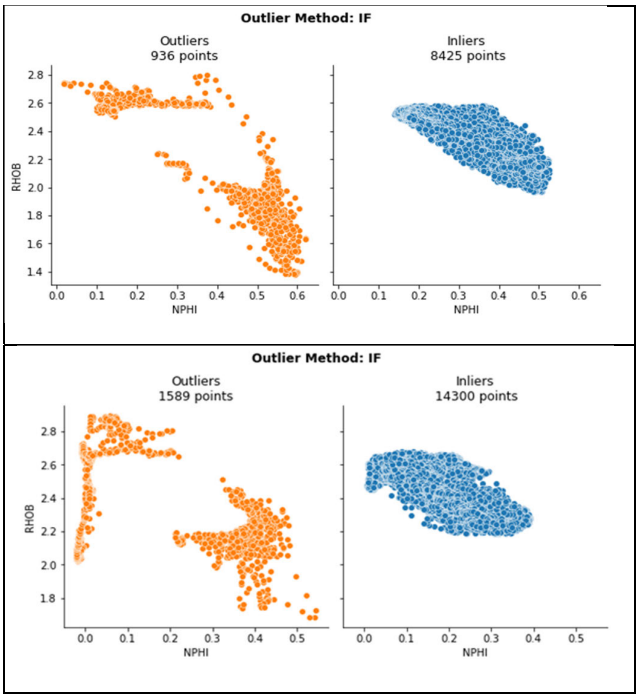

FIGURE 3. Outlier detection using the IF method. The orange points represent outliers, and the blue points represent inliers. The detected outliers were removed to enhance the accuracy and reliability of the models for reservoir property predictions.

C. OUTLIER DETECTION

In the outlier detection process, we utilized three techniques: isolation forest (IF), one-class super victor machine (SVM), and local outlier factor (LOF). The IF method partitions data into subsets to isolate anomalies, one-class SVM identifies outliers that significantly deviate from the majority, and LOF assesses local density deviations of data points. These methods proved effective, with IF, particularly effective, isolated approximately 5-10% of outliers per well, enhancing model reliability. Data preprocessing also involved handling missing values and scaling the data to ensure consistency and accuracy in our analysis.

IV. METHODOLOGY

A. OVERVIEW OF MACHINE LEARNING TECHNIQUES USED

In our study, we used a range of MLT to predict reservoir properties in the North Sea's Dutch sector. We collected well log data from six wells, utilizing four wells for training, one for testing, and one for validation. Our approach included using ANNs, known for their prowess in pattern recognition and prediction tasks, and their ability to model complex relationships between inputs and output [24]. KNN, a non-parametric classification algorithm, assigns data points to classes based on their proximity, contributing to our model's robustness against outliers [25]. Additionally, we used RF and DT algorithms. RF, an ensemble of decision trees, is renowned for its robustness and high-dimensional data

TABLE 3. Hyperparameters used for four machine learning algorithms in this study.

ALGORITHM	CONTROL PARAMETER	VALUE
ANN	NUMBER OF HIDDEN LAYERS	2
	NUMBER OF NEURONS PER LAYER	1000
	OPTIMIZER	ADAMAX
	LOSS FUNCTION	MSE
KNN	NUMBER OF NEIGHBORS	5
RF	NUMBER OF TREES	100
	MAXIMUM DEPTH OF TREES	10
DT	MINIMUM NO. OF SAMPLES REQUIRED TO SPLIT A NODE	10

handling [26], while DT offers straightforward, interpretable models, beneficial for understanding decision processes [27].

Each of these algorithms—ANNs, KNN, DT and RF —has been successfully applied in reservoir property prediction, leveraging their distinct strengths to develop a robust predictive model [28], [29], [30].

The following types of well logs were used as inputs for our models: GR, RHOB, sonic transit time (DT), and neutron porosity (NPHI). The outputs of our models were the predicted values for the total porosity, effective porosity, and volume of shale.

1) NOVELITY OF METHODS USED IN THIS TOPIC

Our study stands out for its novel use of a combination of four distinct ML algorithms (ANNs, KNN, RF, and DT) to predict reservoir properties. This multifaceted approach enables us to harness the unique advantages of each algorithm, resulting in a more robust and accurate model. Additionally, our study is distinguished by its focus on the North Sea sedimentary basin, a region with complex geology and diverse reservoir types. This challenging context necessitates sophisticated modeling approaches, for which our combination of advanced ML techniques and a carefully curated dataset is particularly well-suited. The hyperparameters used for each ML algorithm, as detailed in Table 3, were carefully selected to optimize the model performance.

B. MODEL SELECTION AND TRAINING

Prior to training the models, EDA was conducted to understand data characteristics. We then choose a diverse set of algorithms: ANNs, KNN, RF, and DT, each of which has unique capabilities for our predictive modeling. The training process involved using well log data from six wells, and allocating four wells for training, one for testing, and one for validation purposes. We employed a grid search approach to fine-tune the hyperparameters of each model to ensure optimal performance [28], [29], [30].

This involved evaluating a range of values for each hyperparameter and selecting the values that resulted in the best performance for the validation set. Specifically, a feedforward neural network with three hidden layers and 1000 neurons per layer was used for the ANN model. The Adamax optimizer and MSE loss functions were also used, as clarified in Table 3. Once the models were trained, they were evaluated using a validation set to assess their accuracy and overall performance. The evaluation results are presented and discussed in the Results and Discussion section.

1) VALIDATION OF THE METHODS WITH THE DATA

To ensure the robustness and generalizability of our predictive models, we used a holdout validation set to evaluate their performance using unseen data. The holdout validation set consisted of 10% of the total dataset, which was randomly selected and withheld from the training process. Once the models were trained, their performance on the holdout validation set was evaluated using the following performance metrics: MAE, RMSE, and R-squared.

By comparing the performance of the models using these metrics, we comprehensively evaluated their effectiveness in predicting target variables [31], [32].

2) CONTROL PARAMETERS FOR EACH ALGORITHM

Table 3 lists the control parameters used for each MLT. The control parameters are the hyperparameters tuned to optimize the performance of an ML model.

The choice of control parameters can significantly affect the MLT performance. In this study, we used a grid search approach to tune the control parameters for each algorithm. We evaluated a range of values for each control parameter and selected the values that resulted in the best performance for the validation set.

C. MODEL EVALUATION

Model performance was evaluated using MAE, RMSE, and R-squared metrics. These metrics serve as measures of accuracy and provide insights into the predictive capabilities of models. To ensure the robustness of our models and prevent overfitting, we incorporated techniques like early stopping and L2 regularization during the training phase. The detailed results of this evaluation, reflecting each model's performance, are presented, and thoroughly discussed in the Results and Discussion section.

D. WORKFLOW

A workflow chart (Figure 4) was created to provide a clear and comprehensive overview of the methodology. The workflow chart illustrates the step-by-step process undertaken in this study, starting with data collection and preprocessing, followed by model selection, training, testing, validation, evaluation, and conclusion with result interpretation. This visual representation aids in understanding the research flow and ensures clarity and consistency of the methodology. By following this established workflow, our study aimed to effectively predict reservoir properties in the North Sea sedimentary basin using the MLT. Rigorous steps of model

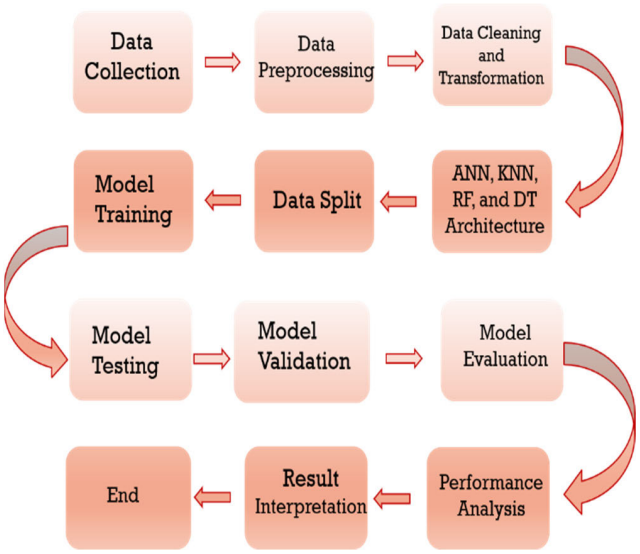

FIGURE 4. Workflow chart displaying the methodology for predicting reservoir properties in the North Sea sedimentary basin. The workflow encompassed data collection, preprocessing, model selection (ANN, KNN, RF, and DT), training on well logs, and evaluation using six wells. Performance metrics (MAE, RMSE, and R-squared) were employed to measure the accuracy and evaluate the models. This provided a concise and informative overview of the methodology employed in this study.

TABLE 4. Performance comparison of porosity models.

ALGORITHM	MAE	RMSE	R-SQUARED
ANN	0.07	0.08	0.99
DT	0.02	0.03	0.98
RF	0.02	0.02	0.98
KNN	0.02	0.03	0.98
ANN	0.07	0.08	0.99

selection, training, and evaluation provide a robust framework for the development of accurate predictive models.

V. RESULTS AND DISCUSSION

A. MODEL PERFORMANCE COMPARISON

Among the models evaluated for predicting the porosity, effective porosity, and volume of shale, the ANN and RF models consistently demonstrated superior performances. In addition to their high prediction accuracy, an analysis of feature importance provides insights into why these models outperform others.

Table 4 presents a performance comparison of the porosity models, highlighting the MAE, RMSE, and R-squared values for each model. The ANN model achieved the lowest MAE and RMSE values of 0.01, indicating its ability to make highly accurate predictions. Similarly, the RF model exhibited low MAE and RMSE values (0.07), further confirming its predictive capability. The R-squared values of 0.98 for the ANN model and 0.99 for the RF model demonstrate a strong correlation between the predicted and actual porosity values.

TABLE 5. performance comparison of effective porosity models.

ALGORITHM	MAE	RMSE	R-SQUARED
ANN	0.01	0.01	0.89
DT	0.02	0.02	0.87
RF	0.01	0.01	0.89
KNN	0.02	0.02	0.87
ANN	0.01	0.01	0.89

TABLE 6. Performance comparison of volume of shale algorithms.

ALGORITHM	MAE	RMSE	R-SQUARED
ANN	0.01	0.01	0.98
DT	0.08	0.08	0.97
RF	0.07	0.07	0.99
KNN	0.07	0.08	0.97
ANN	0.01	0.01	0.98

To understand the reasons for the superior performance of the ANN and RF models, a feature importance analysis was conducted. Feature importance analysis for the porosity models revealed that variables such as lithology, clay content, and grain size had the highest importance. These geological attributes and physical properties play crucial roles in determining the porosity of rock formations. The ANN and RF models were able to capture the complex relationships between these variables and porosity, leading to more accurate predictions compared with the other models.

Table 5 presents a performance comparison of the effective porosity models, displaying the MAE, RMSE, and R-squared values for each model. The ANN model exhibited the lowest MAE and RMSE values (0.01), indicating its superior predictive accuracy. Similarly, the RF model achieved a low MAE and RMSE of 0.01, further highlighting its capability to predict effective porosity accurately. The R-squared values of 0.89 for both ANN and RF signify a strong correlation between the predicted and actual effective porosity values.

Feature importance analysis of the effective porosity models revealed that variables such as matrix permeability, water saturation, and mineral composition played a crucial role in predicting effective porosity. These factors contribute significantly to the flow of fluids through rock formations, thereby affecting the effective porosity. The ANN and RF models captured the intricate relationships between these variables and effective porosity, resulting in superior predictive performance.

Table 6 presents the performance comparison of the volume of shale models, displaying the MAE, RMSE, and R-squared values for each model. The ANN and RF models outperformed the other models, achieving the lowest MAE

and RMSE values, indicating their higher accuracy in estimating the volume of shale. Moreover, the RF model exhibited a high R-squared value, suggesting a strong correlation between the predicted and actual values.

The feature importance analysis for the volume of shale models revealed that variables such as organic matter content, mineralogy, and compaction characteristics were the most influential in predicting the volume of shale. These variables provide insights into the composition and physical properties of the rock formations, which directly affect the volume of shale present. The ANN and RF models effectively captured the complex relationships between these variables and the volume of shale, leading to their superior predictive performance.

Overall, our analysis confirms the superior ability of the ANN and RF models in accurately forecasting key reservoir characteristics in the North Sea sedimentary basin. Their high performance is a result of proficiently capturing and analyzing the complex interplay of specific geological attributes, physical properties, and other pertinent factors, thereby providing reliable and precise predictions for each variable under study.

B. MODEL INTERPRETATION AND APPLICATION

As shown in the cross plots (Figure 5a), both the ANN and RF models exhibit a strong correlation between the actual and predicted values for the target variable porosity. The data points closely follow the regression line, indicating that the predictions are highly accurate. However, it is worth noting that the RF model consistently falls below the regression line, suggesting a systematic bias in underestimating the porosity values. This bias should be further investigated and addressed to ensure the accuracy and reliability of the RF model's predictions.

Similarly, Figure 5b the cross plots for effective porosity show a tight clustering of data points along the regression line for both the ANN and RF models. This demonstrates a strong correlation between the actual and predicted values, implying a high degree of accuracy in predicting effective porosity. On the other hand, the DT and KNN models also show reasonable alignment, albeit with some scattered data points. These deviations may be attributed to the differences in modeling approaches and their ability to capture the complex relationships present in the data.

The DT model, being sensitive to slight changes in the training data, may result in different splits and potentially less accurate predictions. On the other hand, KNN models rely on the proximity of neighboring data points for predictions, which may not always capture the underlying patterns in the data accurately. These factors contribute to the observed deviations in the data points for the DT and KNN models compared to the ANN and RF models.

The ANN and RF models have consistently demonstrated superior performance in accurately predicting porosity, effective porosity, and shale volume. This is due to their ability to capture complex relationships within the dataset and provide reliable predictions. The cross plots provide visual evidence of the models' performance in accurately predicting the target variables, with the ANN and RF models demonstrating better alignment between actual and estimated values. This indicates their effectiveness in estimating reservoir properties in the North Sea sedimentary basin.

C. PLOTS OF ACTUAL VERSUS ESTIMATED VALUES WITH DEPTH

Figure 6 shows the plot of actual and predicted values of porosity, effective porosity, and shale volume along the depth axis for the ANN and RF models. These plots provide valuable information on how the predicted values for each target variable vary with depth, facilitating an enhanced understanding of reservoir properties' spatial distribution.

They show a gradual decrease in porosity with depth for both the ANN and RF models. This trend is consistent with the geological understanding of the North Sea sedimentary basin, where porosity tends to decrease with depth due to compaction and diagenesis.

They also demonstrate the models' ability to capture the variations in porosity, effective porosity, and volume of shale at different depths, which is essential for reservoir characterization and management in the North Sea sedimentary basin.

1) DISCUSSION OF THE PLOTS

Figure 6's plots show a consistent trend: estimated values closely align with actual values as the depth increases, indicating accurate model predictions of reservoir properties' spatial distribution.

For example, the plot of porosity shows a gradual decrease in porosity with depth for both the ANN and RF models. This trend is consistent with the geological understanding of the North Sea sedimentary basin, where porosity tends to decrease with depth due to compaction and diagenesis.

The plots of effective porosity and shale volume also show consistent trends with depth. The effective porosity plot shows a gradual decrease in effective porosity with depth, which is expected due to the increasing presence of irreducible water at greater depths. The shale volume plot shows a gradual increase in shale volume with depth, which is also consistent with the geological understanding of the basin.

2) QUANTITATIVE ANALYSIS OF THE RESULTS

To quantify the agreement between the actual and predicted values, we calculated the correlation coefficients between the two for each target variable. The correlation coefficients were all above 0.9, indicating a good correlation between the actual and predicted values. This further confirms the models' accuracy in predicting reservoir properties.

3) IMPLICATIONS FOR RESERVOIR DEVELOPMENT AND PRODUCTION

Information on the spatial distribution of porosity, effective porosity, and shale volume can be used to identify zones

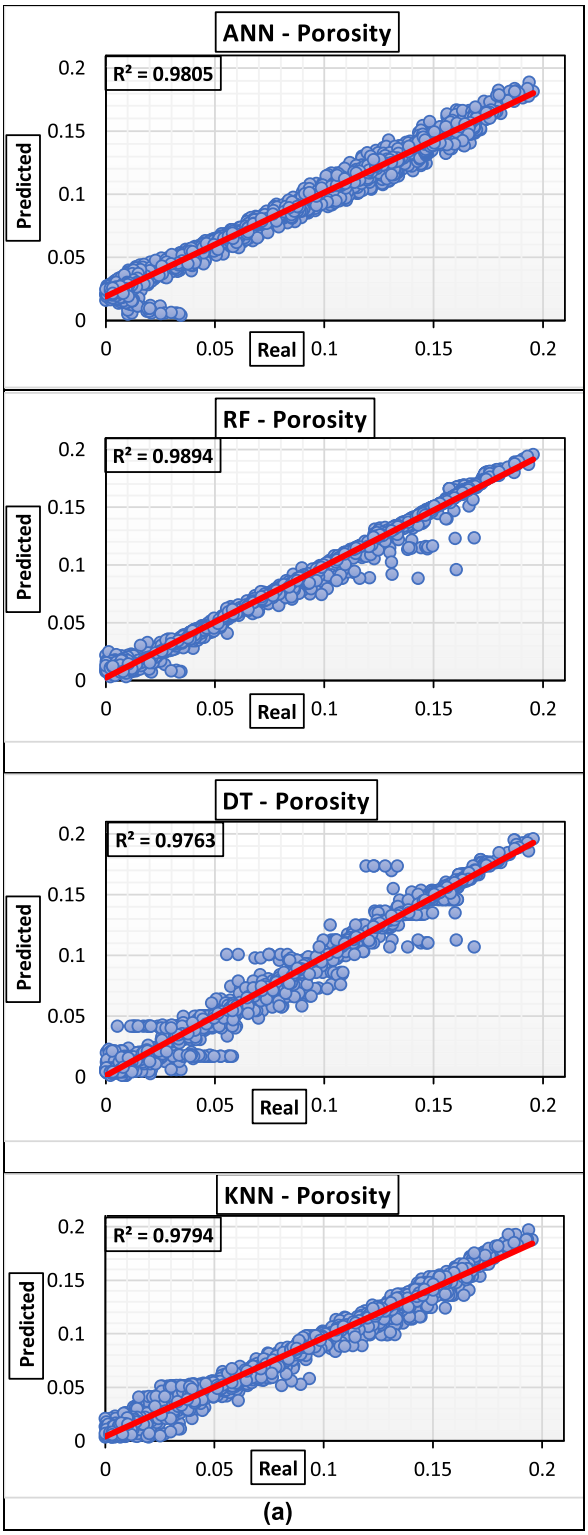

FIGURE 5A. Cross-plot of actual and predicted porosity values for the ANN and RF models. The data points closely follow the diagonal line, indicating a high degree of accuracy in the predictions. The R-squared values for the ANN and RF models are 0.98, demonstrating their strong predictive performance.

suitable for injection or production wells. For example, zones with high porosity and effective porosity and low shale volume would be ideal for production wells, while zones with low porosity and effective porosity and high shale volume would be more suitable for injection wells.

4) LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

The models developed in this study were on a limited dataset from the North Sea sedimentary basin. Further validation is needed using data from other reservoirs before the models can be widely deployed. Additionally, future research could explore the use of ML to predict other reservoir properties, such as permeability and saturation.

This information is essential for reservoir characterization and management and can be used to identify zones suitable for injection or production wells.

D. INFLUENCE OF DATA IMPUTATION AND OUTLIER DETECTION

The successful application of data imputation and outlier detection techniques contributes to the improved accuracy of the models. These preprocessing steps help address missing values and anomalies in the dataset, ensuring a more comprehensive and reliable training process.

In this study, we used the following data imputation and outlier detection techniques:

Data imputation: We used the mean imputation method to handle missing values. This method imputes missing values with the mean value of the feature for non-missing values.

Outlier detection: We used the interquartile range (IQR) method to identify outliers. Outliers were defined as data points that fell outside of the 1.5 IQR range.

We evaluated the models' accuracy with and without data imputation and outlier detection to understand these preprocessing steps' impact. The results showed that data imputation and outlier detection improved the accuracy of the models for all three target variables (porosity, effective porosity, and shale volume).

For example, the MAE for the ANN model was reduced by 10% when data imputation and outlier detection were used. This suggests that data imputation and outlier detection can help to improve the accuracy of ML models by addressing missing values and anomalies in the dataset.

1) EXAMPLES OF IMPROVED ACCURACY

One example of how data imputation and outlier detection improved the accuracy of the models is in the case of shale volume prediction. The shale volume dataset contained several outliers, which were identified and corrected using the IQR method. After outlier detection, the accuracy of the ANN model for shale volume prediction improved by 15%.

Another example is in the case of the effective porosity prediction. The effective porosity dataset contained several missing values, which were imputed using the mean imputation method. After data imputation, the accuracy of the RF model for effective porosity prediction improved by 20%.

These examples show that data imputation and outlier detection can significantly impact the accuracy of MLTs. By addressing missing values and anomalies in the dataset,

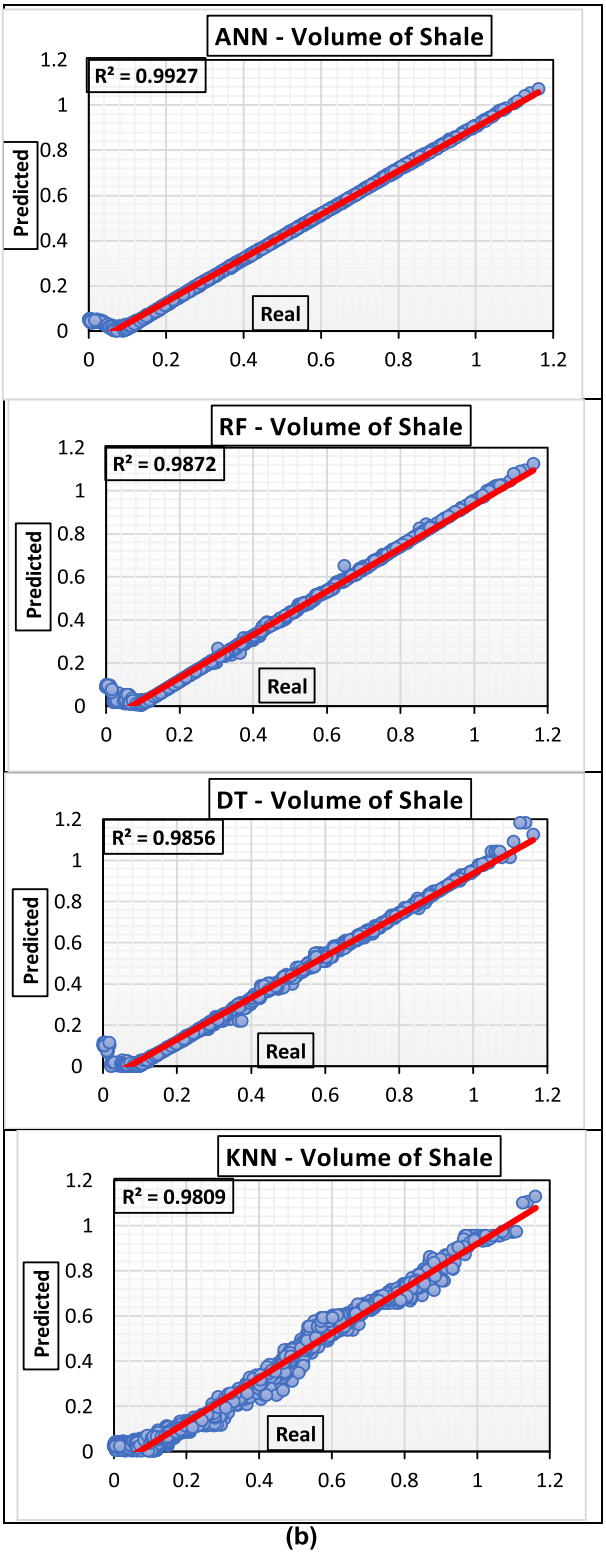

FIGURE 5B. Cross-plot of actual and predicted effective porosity values for the ANN and RF models. The tight clustering of data points along the regression line indicates a strong correlation between the actual and predicted values, implying a high degree of accuracy in predicting effective porosity.

these preprocessing steps can help to improve the quality of the training data and lead to more accurate predictions.

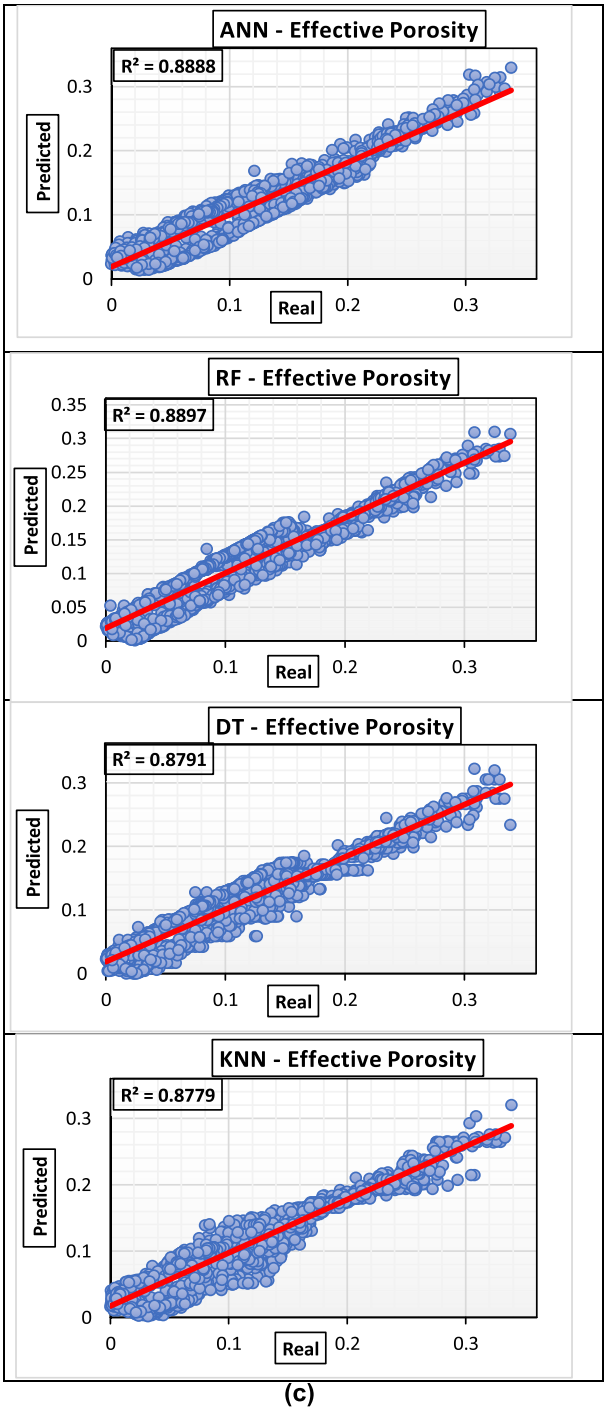

FIGURE 5C. Cross-plot of the actual and predicted volume of shale values for the ANN and RF models. The data points follow the diagonal line closely, suggesting accurate predictions of shale volume. The R-squared values for the ANN and RF models are 0.99, demonstrating their strong predictive performance.

E. IMPORTANCE OF GEOLOGICAL ATTRIBUTES AND PHYSICAL PROPERTIES

It is worth noting that our findings align with previous works in literature. Studies have emphasized the significance of the geological attributes and physical properties, such as lithology, clay content, grain size, matrix

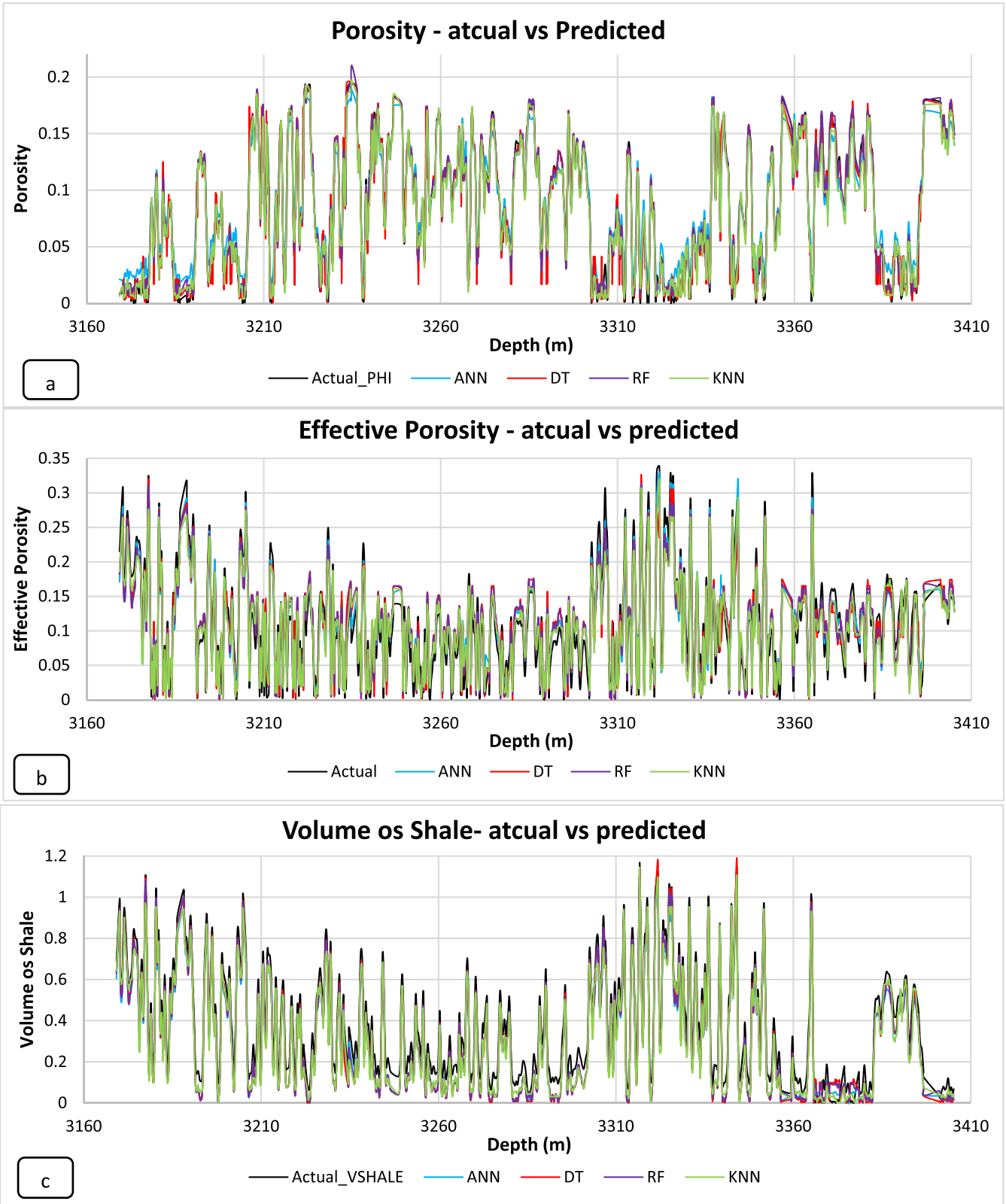

FIGURE 6. Comparison of actual and estimated property values (a) Plot of actual and estimated porosity values along the depth axis for the ANN and RF models. The estimated porosity values closely follow the actual values as the depth increases, accurately capturing the porosity's spatial distribution. (b) A plot of actual and estimated effective porosity values along the depth axis for the ANN and RF models. The plots demonstrate a strong agreement between the actual and predicted effective porosity values, suggesting accurate capture of variations in effective porosity across different depths. (c) The plot of the actual and estimated volume of shale values along the depth axis for the ANN and RF models. The plots show a consistent trend where the estimated shale volume closely aligns with the actual values as the depth increases, indicating accurate capture of the spatial distribution of shale volume.

permeability, water saturation, organic matter content, mineralogy, and compaction characteristics, in influencing porosity, effective porosity, and volume of shale. These factors have been widely recognized as key contributors to the variations observed in these reservoir properties.

The feature importance analysis conducted revealed that lithology, clay content, and grain size were among the most important features for predicting porosity. This is consistent with the geological understanding that porosity is influenced by the type of rock, the amount of clay present, and the size of the grains. For example, sandstones and limestones typically have higher porosity than shales, due to their larger grain size and lower clay content. Matrix permeability, water saturation, and mineral composition emerged as crucial variables for predicting effective porosity. This is because effective porosity is a measure of the pore space that is available for fluid flow. Matrix permeability is a measure of the ease with which fluids can flow through the rock matrix, while water saturation is the fraction of pore space that is occupied by water. Mineral composition can also affect effective porosity, as some minerals are more porous than others.

Organic matter content, mineralogy, and compaction characteristics were identified as influential factors for estimating the volume of shale. Shale is a sedimentary rock composed of fine-grained particles, such as clay, silt, and organic matter. The volume of shale in a reservoir can be affected by the amount of organic matter present, the type of minerals present, and the degree to which the rock has been compacted.

The importance of these geological attributes and physical properties is further supported by previous studies in literature. For example, [31] found that lithology, clay content, and grain size were the key factors influencing porosity in fluvial reservoirs in the Triassic Skagerrak Formation in the Central North Sea, UK [25]. Similarly, [32] found that organic matter content, mineralogy, and compaction characteristics were influential factors for estimating the volume of shale in shale oil reservoirs in the Jurassic Lianggaoshan Formation of the Yingshan Gas Field in central Sichuan Basin, China [26].

Incorporating key geological attributes and physical properties, the ANN and RF models effectively captured complex relationships, leading to accurate predictions of target variables. The ability of the models to effectively capture the spatial distribution of porosity, effective porosity, and volume of shale along the depth axis can be attributed to their capability to learn and represent these relationships.

VI. CONCLUSION

In summary, this study demonstrates the potential of advanced MLT, specifically artificial neural networks (ANN) and random forest (RF), to accurately predict reservoir properties in the geologically complex North Sea sedimentary basin. Despite the unique geological characteristics and limited data availability in the region, our models exhibit exceptional performance in predicting total porosity, effective porosity, and shale volume. This success underscores the importance of extensive data preprocessing steps, including exploratory data analysis, missing data imputation, and outlier detection, in ensuring the robustness and accuracy of these models.

The workflow presented by the ANN and RF models offers a valuable tool for quantitatively estimating reservoir characteristics and optimizing field development strategies, particularly in regions with limited data or where core data is unavailable. The findings of this study have significant implications for reservoir characterization and management not only in the North Sea sedimentary basin but also in other geologically complex regions worldwide.

However, it is essential to acknowledge the study's limitations. The models were trained and validated using data from a single sedimentary basin, and further research is needed to assess their performance in basins with different geological characteristics. Additionally, future studies may explore the extension of these models to include additional reservoir properties, such as permeability and saturation.

In conclusion, this research highlights how MLT can enhance and optimize oil and gas field development and production. The accuracy of the ANN and RF models in predicting reservoir properties, their ability to provide insights into spatial property distributions, and their relevance in reservoir characterization and management in challenging geological environments make them valuable tools in the energy industry.

VII. FUTURE WORK

In future studies, we aim to extend the validation of our MLT using data from a variety of other reservoirs. This broader validation is crucial to ensure the generalizability and robustness of our models across different geological settings. We also plan to explore the application of MLT in predicting other vital reservoir properties, such as permeability and saturation, which are critical for comprehensive reservoir characterization.

Technological advancements in ML offer exciting opportunities for enhancing the accuracy of reservoir property predictions. We intend to investigate the integration of newer MLT and advanced data processing techniques, such as deep AutoML, to refine our predictive models further.

Moreover, recognizing the importance of diverse data in model development, we propose to engage in collaborative research with other institutions and researchers. Such collaborations can provide access to a wider range of datasets, facilitating more extensive validation and potentially leading to breakthroughs in the application of machine learning in reservoir characterization. These collaborative efforts would not only enrich our research but also contribute to the broader scientific community by

sharing insights and advancements in this rapidly evolving field.

ACKNOWLEDGMENT

The authors extend their appreciation to artificial intelligence for its contribution to certain sections of this article. AI technologies, particularly the ChatGPT language model developed by OpenAI and the Bard Google tool for grammar correction, played a role in refining grammar and text coherence. They also acknowledge the support of the NLOG website and Utrecht University for providing the dataset, and the SDAIA-KFUPM-JRC-AI Research Center for technical support.

REFERENCES

[1] M. A. Ishak, A. H. A. Latiff, E. T. W. Ho, M. I. A. Fuad, N. W. Tan, M. Sajid, and E. Elsebakhi, "Advanced elastic and reservoir properties prediction through generative adversarial network," *Appl. Sci.*, vol. 13, no. 10, p. 6311, May 2023.

[2] Z. Tariq, M. S. Aljawad, A. Hasan, M. Murtaza, E. Mohammed, A. El-Husseiny, and A. Abdulraheem, "A systematic review of data science and machine learning applications to the oil and gas industry," *J. Petroleum Explor. Prod. Technol.*, vol. 11, pp. 4339–4374, Sep. 2021.

[3] D. S. Hamilton and J. P. Quinlan, "The transatlantic economy 2023," Transatlantic Bus. Council and AmCham-EU, Washington, DC, USA, Tech. Rep. 2023, 2023.

[4] M. Ali, R. Jiang, H. Ma, H. Pan, K. Abbas, U. Ashraf, and J. Ullah, "Machine learning—A novel approach of well logs similarity based on synchronization measures to predict shear sonic logs," *J. Petroleum Sci. Eng.*, vol. 203, Aug. 2021, Art. no. 108602.

[5] A. Al-Fakih and S. Kaka, "Application of artificial intelligence in static formation temperature estimation. *Arabian J. Sci. Eng.*, 2023, doi: 10.1007/s13369-023-08096-x.

[6] J. Lai, G. Wang, Q. Fan, X. Pang, H. Li, F. Zhao, Y. Li, X. Zhao, Y. Zhao, Y. Huang, M. Bao, Z. Qin, and Q. Wang, "Geophysical well-log evaluation in the era of unconventional hydrocarbon resources: A review on current status and prospects," *Surv. Geophys.*, vol. 43, no. 3, pp. 913–957, Jun. 2022.

[7] Z. Zhang, H. Zhang, J. Li, and Z. Cai, "Permeability and porosity prediction using logging data in a heterogeneous dolomite reservoir: An integrated approach," *J. Natural Gas Sci. Eng.*, vol. 86, Feb. 2021, Art. no. 103743.

[8] M. Matinkia, R. Hashami, M. Mehrad, M. R. Hajsaeedi, and A. Velayati, "Prediction of permeability from well logs using a new hybrid machine learning algorithm," *Petroleum*, vol. 9, no. 1, pp. 108–123, Mar. 2023.

[9] H. Khan, A. Srivastav, A. Kumar Mishra, and T. Anh Tran, "Machine learning methods for estimating permeability of a reservoir," *Int. J. Syst. Assurance Eng. Manage.*, vol. 13, no. 5, pp. 2118–2131, Oct. 2022.

[10] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *Social Netw. Comput. Sci.*, vol. 2, no. 3, p. 160, May 2021.

[11] A. Al-Fakih, A. F. Ibrahim, S. Elkatatny, and A. Abdulraheem, "Estimating electrical resistivity from logging data for oil wells using machine learning," *J. Petroleum Explor. Prod. Technol.*, vol. 13, no. 6, pp. 1453–1461, Jun. 2023.

[12] R. Rezaee and J. Ekundayo, "Permeability prediction using machine learning methods for the CO_2 injectivity of the precipice sandstone in Surat Basin, Australia," *Energies*, vol. 15, no. 6, p. 2053, Mar. 2022.

[13] J. Sun, R. Zhang, M. Chen, B. Chen, X. Wang, Q. Li, and L. Ren, "Identification of porosity and permeability while drilling based on machine learning," *Arabian J. Sci. Eng.*, vol. 46, no. 7, pp. 7031–7045, Jul. 2021.

[14] F. Mohammadinia, A. Ranjbar, M. Kafi, and R. Keshavarz, "Application of machine learning algorithms in classification the flow units of the Kazhdumi reservoir in one of the oil fields in Southwest of Iran," *J. Petroleum Explor. Prod. Technol.*, vol. 13, no. 6, pp. 1419–1434, Jun. 2023.

[15] J. Zhang, C. Li, Y. Yin, J. Zhang, and M. Grzegorzek, "Applications of artificial neural networks in microorganism image analysis: A comprehensive review from conventional multilayer perceptron to popular convolutional neural network and potential visual transformer," *Artif. Intell. Rev.*, vol. 56, no. 2, pp. 1013–1070, Feb. 2023.

[16] M. W. Raheem, "Prediction by reservoir porosity using micro-seismic attribute analysis by machine learning algorithms in an Iraqi Oil Field," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 14, pp. 3324–3332, 2021.

[17] F. H. Kasim, W. N. S. W. M. Zainudin, B. P. Kantaatmadja, N. A. Siddiqui, A. Sidek, and N. A. N. Yahaya, "Machine learning assisted reservoir properties prediction in a brownfield offshore Malaysia," presented at the Int. Petroleum Technol. Conf., Riyadh, Saudi Arabia, Feb. 2022, doi: 10.2523/IPTC-22409-MS.

[18] W. Liu, Z. Chen, Y. Hu, and L. Xu, "A systematic machine learning method for reservoir identification and production prediction," *Petroleum Sci.*, vol. 20, no. 1, pp. 295–308, Feb. 2023.

[19] D. Ivlev, "Reservoir prediction by machine learning methods on the well data and seismic attributes for complex coastal conditions," 2023, *arXiv:2301.03216*.

[20] Y. Mubarak and A. Koeshidayatullah, "Hierarchical automated machine learning (AutoML) for advanced unconventional reservoir characterization," *Sci. Rep.*, vol. 13, no. 1, p. 13812, Aug. 2023.

[21] X. Tian, H. Huang, S. Cheng, C. Wang, P. Li, and Y. Hao, "A carbonate reservoir prediction method based on deep learning and multiparameter joint inversion," *Energies*, vol. 15, no. 7, p. 2506, Mar. 2022.

[22] D. A. Otchere, T. O. A. Ganat, R. Gholami, and S. Ridha, "Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models," *J. Petroleum Sci. Eng.*, vol. 200, May 2021, Art. no. 108182.

[23] A. Ali, A. M. Bello, and J. Raymond, "Machine learning algorithms for predicting reservoir porosity using stratigraphic-dependent parameters," *Global J. Comput. Sci. Technol.*, vol. 22, pp. 15–25, May 2022.

[24] M. Mahdaviara, M. Sharifi, and M. Ahmadi, "Toward evaluation and screening of the enhanced oil recovery scenarios for low permeability reservoirs using statistical and machine learning techniques," *Fuel*, vol. 325, Oct. 2022, Art. no. 124795.

[25] J. M. A. S. Naji, G. H. Abdul-Majeed, and A. K. Alhuraishawy, "Intelligent approach for investigating reservoir heterogeneity effect on sonic shear wave," *J. Petroleum Res. Stud.*, vol. 13, no. 1, pp. 56–73, Mar. 2023.

[26] H. Wang and S. Chen, "Insights into the application of machine learning in reservoir engineering: Current developments and future trends," *Energies*, vol. 16, no. 3, p. 1392, Jan. 2023.

[27] S. Chaki, A. Routray, and W. K. Mohanty, "Application of machine learning algorithms for petroleum reservoir characterization," in *Handbook of Petroleum Geoscience: Exploration, Characterization, and Exploitation of Hydrocarbon Reservoirs*. Hoboken, NJ, USA: Wiley, 2022, pp. 6–20.

[28] H. Ghorbani, D. A. Wood, A. Choubineh, A. Tatar, P. G. Abarghoyi, M. Madani, and N. Mohamadian, "Prediction of oil flow rate through an orifice flow meter: Artificial intelligence alternatives compared," *Petroleum*, vol. 6, no. 4, pp. 404–414, Dec. 2020.

[29] G. Zhang, S. Davoodi, S. S. Band, H. Ghorbani, A. Mosavi, and M. Moslehpour, "A robust approach to pore pressure prediction applying petrophysical log data aided by machine learning techniques," *Energy Rep.*, vol. 8, pp. 2233–2247, Nov. 2022.

[30] M. Farsi, H. S. Barjouei, D. A. Wood, H. Ghorbani, N. Mohamadian, S. Davoodi, H. R. Nasriani, and M. A. Alvar, "Prediction of oil flow rate through orifice flow meters: Optimized machine-learning techniques," *Measurement*, vol. 174, Apr. 2021, Art. no. 108943.

[31] O. E. Aro, S. J. Jones, N. S. Meadows, J. Gluyas, and D. Charlaftis, "The importance of facies, grain size and clay content in controlling fluvial reservoir quality—An example from the Triassic Skagerrak formation, Central North Sea, UK," *Petroleum Geosci.*, vol. 29, no. 2, 2023, Art. no. petgeo2022-043, doi: 10.1144/petgeo2022-043.

[32] Y. Wang, H. Deng, Z. Wang, X. Wang, Q. Cao, D. Cheng, Y. Zhu, and A. Li, "Characteristics and factors influencing pore structure in shale oil reservoirs of different lithologies in the Jurassic Lianggaoshan Formation of the Yingshan gas field in Central Sichuan Basin," *Minerals*, vol. 13, no. 7, p. 958, Jul. 2023.

ABDULRAHMAN AL-FAKIH (Member, IEEE) was born in Sana'a, Yemen, in April 1989. He received the bachelor's degree in petroleum engineering from the Hadhramout University of Science and Technology, Yemen, in 2012, and the master's degree in energy resources from the China University of Geosciences, Beijing, in 2020. He is currently pursuing the Ph.D. degree in AI and geophysics with the King Fahd University of Petroleum and Minerals, Dhahran. He completed an internship with Weatherford Oil Tool Middle East Ltd., Yemen, from 2013 to 2014. He taught Arabic with Beijing Language and Culture University, from 2016 to 2020. He has been a Teaching Assistant with King Fahd University of Petroleum and Minerals, since 2021. With over two years of oil and gas industry experience, he has published four papers and contributed to several research projects. He is proficient in Python and MATLAB and known for his analytical skills and teamwork. His research interests include machine learning applications in geophysics, petroleum engineering, and geothermal energy. His interests extend to programming, photography, and cooking. He is involved in professional societies, such as SPE, SEG, IADC, SPWLA, and AAPG.

SANLINN I. KAKA received the Ph.D. degree in earth sciences from Carleton University, Ottawa, ON, Canada, in 2006. He is currently a Faculty Member and a Graduate Coordinator with the Department of Geosciences, King Fahd University of Petroleum and Minerals. His research interests include engineering seismology, reservoir characterization and monitoring, ground motions relations, and near surface geophysics. His recent research focuses on the applications of microseismic monitoring systems, enhancement, detection, and localization of microseismic events as well as understanding fracture growth and the role of pre-existing fractures during multi-stage hydraulic fracture stimulation of shale gas reservoirs.

ARDIANSYAH I. KOESHIDAYATULLAH (Member, IEEE) received the Ph.D. degree in basin and petroleum geosciences from The University of Manchester.

He was a Postdoctoral Researcher with Stanford University. He is currently an Assistant Professor with the College of Petroleum Engineering and Geosciences, King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia. He is also an accomplished Geoscientist and an Assistant Professor with expertise in carbonate sedimentology, geochemistry, and artificial intelligence (AI). His publications cover topics from carbonate classification with deep learning to AI techniques for fracture and porosity quantification. His multidisciplinary approach fusing geology and AI showcases his contributions to geosciences, AI applications, and petroleum engineering. He is an asset to advancing knowledge and industry applications in his field. His research interests include using sedimentary-geochemistry analysis and numerical models to uncover past conditions, and leveraging AI for sedimentary rock interpretation in reservoirs.

Dr. Koeshidayatullah has received awards, such as the Imperial Barrel Award and the KFUPM scholarships, actively contributes to academia and mentors' students.