Mrs.K.Tulya Sree Simla, 2026, 14:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

Boosting Ensemble Machine Learning Approach for Porosity Prediction in Carbon Dioxide Storage Reservoirs

Mrs.K.Tulya Sree Simla [1], Manne Namratha Sai [2], Mantri Bala Subrahmanyam [3], Ompolu Janu Priyanka [4], Korubilli Manoj Kumar [5], Garaga Manikyam [6]

[1] Assistant Professor, Department of CSE (Data Science) In Pragati Engineering College, Surampalem, Andhra Pradesh, India,

[2,3,4,5,6] UG Students Department of CSE (Data Science) In Pragati Engineering College, Surampalem, Andhra Pradesh, India.

Abstract- Accurate estimation of reservoir porosity is a critical factor in evaluating geological formations for carbon dioxide (CO_2) storage in carbon capture and storage (CCS) projects. Porosity directly influences the storage capacity and injectivity of subsurface reservoirs, making its accurate prediction essential for effective CO_2 sequestration planning. Traditional porosity estimation methods based on core analysis are reliable but often expensive, time-consuming, and limited in spatial coverage. With the increasing availability of well-log data, machine learning techniques provide an efficient data-driven alternative for predicting reservoir properties. This study proposes a machine learning–based framework for porosity prediction using boosting ensemble algorithms to support CO_2 storage assessment. Well-log data collected from the Mena Murtee-1 well in the Darling Basin, Australia, are used as input features, while laboratory-corrected porosity values serve as the target variable. Data preprocessing techniques are applied to remove noise, handle missing values, and eliminate multicollinearity among input parameters. Ensemble boosting algorithms including AdaBoost Regression, Gradient Boost Regression, and Extreme Gradient Boost Regression (XGBoost) are implemented and evaluated using standard statistical performance metrics. Experimental results demonstrate that boosting ensemble algorithms effectively capture complex non-linear relationships between well-log parameters and porosity values. Among the evaluated models, Extreme Gradient Boost Regression achieves the highest prediction accuracy and provides reliable porosity estimates for subsurface formations. The proposed framework enhances reservoir characterization accuracy and supports efficient evaluation of geological formations for carbon dioxide storage.

INDEX TERMS: Porosity Prediction, Carbon Capture and Storage, Boosting Ensemble Algorithms, Machine Learning, Reservoir Characterization, XGBoost, Well-Log Data, CO_2 Storage Assessment.

I. INTRODUCTION

The increasing concerns regarding climate change and global warming have intensified the need for sustainable technologies that reduce carbon dioxide (CO_2) emissions. Carbon Capture and Storage (CCS) has emerged as one of the most promising solutions for mitigating greenhouse gas emissions by capturing CO_2 from industrial sources and storing it in deep geological formations. The effectiveness of CCS projects largely depends on the accurate assessment of subsurface storage capacity and reservoir characteristics. Among various reservoir properties, porosity plays a critical role because it directly determines the amount of CO_2 that can be stored within a geological formation.

Traditionally, porosity estimation is performed using laboratory core analysis. Core samples extracted during drilling are analysed to determine the physical properties of reservoir rocks. Although this method provides reliable and accurate results, it is expensive, time-consuming, and often limited by the availability and quality of core samples. In many exploration scenarios, only a small number of core samples are available, which restricts the ability to obtain continuous porosity information across the entire reservoir. As a result, alternative techniques that can efficiently estimate porosity using available subsurface data have gained increasing attention in reservoir characterization studies.

With the advancement of modern drilling technologies, large volumes of well-log data are continuously generated during exploration and reservoir assessment processes. Well logs record various physical and geological properties of subsurface formations, such as gamma ray, neutron porosity, sonic velocity, and resistivity measurements. These datasets provide valuable information for understanding the geological structure and reservoir characteristics of subsurface formations. However, due to the high dimensionality and complex relationships among well-log parameters, traditional statistical approaches often struggle to accurately capture the underlying patterns required for reliable porosity prediction.

Machine learning (ML) techniques have become powerful tools for analyzing complex datasets and identifying hidden relationships between input variables and target outputs. In reservoir engineering and geoscience applications, ML models have been widely used for predicting petrophysical properties such as porosity, permeability, and lithology classification. These techniques are capable of learning non-linear relationships from large datasets and can significantly improve prediction accuracy compared to traditional empirical models.

Among various machine learning approaches, ensemble learning methods have attracted considerable attention due to their ability to improve prediction performance by combining multiple learning models. Boosting algorithms, a subset of ensemble learning techniques, iteratively build strong predictive models by combining several weak learners. Popular boosting algorithms such as AdaBoost, Gradient Boosting, and Extreme Gradient Boosting (XGBoost) have demonstrated strong performance in regression and classification problems across many scientific and engineering domains.

Motivated by these advantages, this study proposes a boosting ensemble machine learning framework for predicting reservoir porosity using well-log data to support carbon dioxide storage assessment. The proposed approach utilizes well-log measurements obtained from the Mena Murtee-1 well located in the Darling Basin, Australia. Multiple boosting algorithms are implemented and evaluated to identify the most accurate model for predicting porosity in sandstone-dominated formations. The primary objective of this research is to improve the accuracy and efficiency of porosity estimation while supporting data-driven decision-making in carbon capture and storage projects.

The remainder of this paper is organized as follows. Section II reviews previous studies related to porosity prediction and machine learning techniques in reservoir characterization. Section III presents the analysis of the existing system and the proposed methodology. Section IV describes the system

architecture and design framework. Section V outlines the implementation modules of the proposed model. Section VI discusses the experimental results and performance evaluation. Finally, Section VII summarizes the conclusions and potential future research directions.

II. LITERATURE SURVEY

In recent years, machine learning and data-driven techniques have been widely applied in geoscience and reservoir engineering to improve the prediction of subsurface properties. With the growing interest in carbon capture and storage (CCS) technologies, accurate estimation of reservoir characteristics such as porosity and permeability has become increasingly important for evaluating geological formations suitable for carbon dioxide (CO_2) storage. Traditional analytical methods often struggle to handle large-scale well-log datasets, which has motivated researchers to adopt advanced computational techniques for reservoir characterization [3], [4].

Several studies have explored the use of machine learning algorithms for predicting petrophysical properties from well-log data. Researchers have applied algorithms such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees, and Random Forest to estimate reservoir parameters. These models have demonstrated the ability to capture complex non-linear relationships between well-log measurements and reservoir properties, providing more accurate predictions compared to traditional empirical approaches. However, the performance of these models often depends heavily on data preprocessing and feature engineering techniques used during model development.

To improve prediction accuracy, ensemble learning methods have gained significant attention in recent years. Ensemble techniques combine multiple weak learning models to create a stronger predictive system. Methods such as bagging, boosting, and stacking have been successfully used to enhance model robustness and reduce prediction errors. Boosting algorithms, in particular, have demonstrated superior performance in many regression and classification tasks because they iteratively focus on correcting the errors of previously trained models.

Among boosting algorithms, AdaBoost, Gradient Boosting, and Extreme Gradient Boosting (XGBoost) are widely recognized for their effectiveness in handling complex datasets. These algorithms have been successfully applied in various engineering and scientific applications, including energy forecasting, reservoir characterization, and geological property estimation. The ability of boosting algorithms to handle high-dimensional datasets and model non-linear relationships makes them particularly suitable for predicting reservoir porosity using well-log data. Recent studies have also highlighted the importance of data preprocessing and feature selection in improving machine learning model performance. Well-log datasets often contain missing values, noise, and highly correlated features that can negatively impact model accuracy. Techniques such as data cleaning, normalization, and multicollinearity analysis are commonly applied to ensure reliable predictions. In addition, visualization tools such as correlation matrices and feature importance analysis are frequently used to identify the most influential input variables affecting porosity prediction.

Despite the progress made in applying machine learning techniques to reservoir characterization, several challenges remain. Geological datasets are often complex and heterogeneous, making it difficult for single machine learning models to achieve consistent performance across different formations. Furthermore, limited studies have specifically investigated the application of boosting ensemble algorithms for porosity prediction in the context of carbon dioxide storage assessment.
Therefore, there is a need for robust machine learning frameworks that can effectively handle well-log data complexity while providing accurate and reliable porosity predictions. By leveraging boosting ensemble algorithms and well-log datasets, the proposed study aims to develop an efficient machine learning approach for reservoir porosity prediction, supporting more reliable carbon dioxide storage assessment and subsurface characterization.

III. SYSTEM ANALYSIS

A. Existing System

In the existing system, porosity estimation for carbon dioxide (CO_2) storage assessment is mainly performed using traditional geological and petrophysical methods. One of the most commonly used techniques is core sample analysis, where rock samples obtained during drilling operations are tested in laboratories to determine reservoir properties such as porosity and permeability. Although this method provides accurate and reliable measurements, it is expensive, time-consuming, and limited to specific depths where core samples are available. As a result, it cannot provide continuous porosity estimation across the entire reservoir formation.

In addition to laboratory core analysis, empirical and statistical models are often used to estimate porosity from well-log data. These methods rely on predefined mathematical relationships between well-log parameters and porosity values. Commonly used well logs include gamma ray, neutron porosity, sonic velocity, and resistivity logs. While these approaches are faster than laboratory analysis, they often assume simplified linear relationships between input parameters and reservoir properties. Such assumptions may not accurately represent the complex geological characteristics of subsurface formations.

To overcome these limitations, machine learning techniques have been increasingly applied for predicting reservoir properties using well-log data. Conventional machine learning models such as Linear Regression, Decision Trees, Support Vector Machines (SVM), and Artificial Neural Networks have been used to estimate porosity values from well-log measurements. These models can identify complex relationships between input variables and output parameters, providing improved prediction performance compared to traditional empirical methods.

Furthermore, ensemble learning techniques have been introduced to enhance prediction accuracy and model robustness. Algorithms such as Random Forest and Gradient Boosting combine multiple weak learning models to produce more reliable predictions and reduce the risk of overfitting. These ensemble models have shown promising results in various reservoir characterization studies and geological property prediction tasks.

Recent advancements in data-driven reservoir analysis have also enabled the use of large well-log datasets for predicting subsurface properties. Modern drilling operations generate large volumes of well-log data that can be analysed using machine learning techniques to improve reservoir characterization. However, many existing predictive models still struggle to handle complex geological relationships and large-scale datasets efficiently. Additionally, the prediction accuracy of traditional models may vary significantly depending on the quality of input data and preprocessing techniques used.

Limitations Of Existing System

- Traditional core analysis methods require laboratory testing, which is expensive, time-consuming, and limited to specific sample locations within the reservoir.
- Core samples provide only discrete measurements and cannot represent continuous porosity variations across the entire geological formation.
- Empirical and statistical models often assume simplified linear relationships between well-log parameters and porosity values, which may not accurately capture complex subsurface geological conditions.
- Well-log datasets may contain missing values, noise, and highly correlated features that negatively affect prediction accuracy.
- Conventional machine learning models may suffer from overfitting and reduced prediction reliability when dealing with high-dimensional well-log datasets.
- Many existing approaches lack efficient ensemble learning frameworks capable of improving prediction accuracy and robustness for reservoir characterization tasks.

B. Proposed System

This section presents the proposed machine learning framework for predicting reservoir porosity to support carbon dioxide (CO_2) storage assessment. The proposed system utilizes well-log data collected from geological formations and applies boosting ensemble machine learning algorithms to improve the accuracy of porosity prediction.

In the proposed approach, well-log measurements such as gamma ray, neutron porosity, sonic velocity, and resistivity logs are used as input features. Laboratory-corrected porosity values serve as the target output for training machine learning models. Before model development, several data preprocessing techniques are applied to improve data quality and ensure reliable predictions. These steps include data cleaning, removal of missing values, and multicollinearity analysis to eliminate highly correlated input variables.

The proposed framework employs boosting ensemble regression algorithms, including AdaBoost Regression, Gradient Boost Regression, and Extreme Gradient Boost Regression (XGBoost). These algorithms work by combining multiple weak learners in a sequential manner, where each new model focuses on correcting the prediction errors of previous models. This iterative learning process enables the system to capture complex non-linear relationships between well-log features and porosity values.

Model performance is evaluated using statistical metrics such as coefficient of determination (R^2), mean squared error (MSE), and mean absolute error (MAE). By comparing the performance of different boosting algorithms, the most accurate model for predicting reservoir porosity is identified.

The proposed machine learning framework provides a cost-effective and efficient alternative to traditional porosity estimation methods. By utilizing well-log data and advanced ensemble learning techniques, the system improves prediction accuracy and supports reliable subsurface characterization for carbon dioxide storage assessment.

IV. SYSTEM DESIGN

System Architecture

Below diagram depicts the whole system architecture.

Fig 1. Methodology followed for proposed model

V. SYSTEM IMPLEMENTATION

Modules

This section describes the implementation modules of the proposed machine learning framework developed for predicting reservoir porosity for carbon dioxide (CO_2) storage assessment. The system follows a structured pipeline consisting of data acquisition, preprocessing, feature selection, machine learning model training, and performance evaluation. This modular architecture improves prediction accuracy, system reliability, and computational efficiency when analysing well-log datasets for geological reservoir characterization [3], [4].

A. Data Collection Module

The Data Collection Module gathers geological well-log data obtained from subsurface reservoir formations. Well-log measurements are recorded during drilling operations and provide valuable

information about the physical properties of underground rock formations.

The dataset used in this study contains several well-log parameters commonly used for reservoir characterization, including:

- Gamma Ray Log (GR)
- Neutron Porosity Log (NPHI)
- Density Log (RHOB)
- Sonic Log (DT)
- Resistivity Log (RT)

These parameters provide indirect measurements of rock properties and are widely used to estimate reservoir porosity and permeability. The dataset includes both input well-log measurements and corresponding porosity values obtained through laboratory core analysis.

The collected data are stored in a structured format and transferred to the preprocessing module for further analysis and cleaning. The availability of large well-log datasets enables machine learning models to identify complex relationships between geological parameters and reservoir properties [5], [7].

B. Data Preprocessing Module

The Data Preprocessing Module improves the quality of the well-log dataset before it is used for machine learning model training. Geological datasets often contain missing values, noise, and redundant features that can negatively affect prediction accuracy if not properly handled.

The preprocessing stage includes the following steps:

1) Missing Value Handling

Missing values in well-log measurements may occur due to sensor malfunction or incomplete data acquisition during drilling operations. These missing values are handled using appropriate imputation techniques to ensure dataset completeness and prevent information loss.

2) Data Cleaning and Noise Removal

Outliers and noisy data points are removed from the dataset to ensure accurate representation of geological characteristics. Cleaning the dataset improves model stability and prediction reliability.

3) Data Normalization

Feature scaling and normalization techniques are applied to ensure that all well-log variables operate within a consistent numerical range. This step prevents model bias caused by variables with larger numerical magnitudes.

These preprocessing techniques enhance data quality and improve the robustness of machine learning models used for reservoir property prediction [6], [8].

C. Feature Selection Module

Reservoir datasets often contain multiple well-log parameters, some of which may be redundant or weakly correlated with the target variable. High-dimensional datasets increase computational complexity and may reduce prediction efficiency.

Therefore, a Feature Selection Module is implemented to identify the most relevant well-log parameters affecting porosity prediction.

Feature importance is evaluated using machine learning-based feature ranking methods that determine the contribution of each input variable to the prediction outcome. This process helps identify the most influential geological parameters affecting reservoir porosity.

By selecting only the most significant features, the framework reduces dataset dimensionality, decreases model training time, and improves prediction accuracy. Feature selection also improves interpretability by identifying the key well-log parameters influencing porosity estimation [9], [10].

D. Machine Learning Training Module

The Machine Learning Training Module develops regression models to predict reservoir porosity using well-log data. Several machine learning algorithms

are implemented and evaluated to determine the most effective predictive model.

The algorithms used in this study include:

- AdaBoost Regression
- Gradient Boosting Regression
- Extreme Gradient Boosting (XGBoost)

Boosting algorithms are ensemble learning methods that combine multiple weak learning models to create a strong predictive model. Each new model focuses on correcting the prediction errors made by previous models, thereby improving overall prediction accuracy.

These algorithms are particularly effective in modelling complex non-linear relationships between geological variables and reservoir properties. Boosting models also provide improved prediction stability and reduced overfitting when compared to traditional regression models [11], [12].

E. Prediction and Evaluation Module

The Prediction and Evaluation Module generates the final porosity prediction results and evaluates the performance of the trained machine learning models.

The output of the system includes:

- Predicted reservoir porosity values
- Model prediction accuracy
- Feature importance ranking

To evaluate model performance, several statistical evaluation metrics are used:

- Coefficient of Determination (R^2)
- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)

These metrics provide a comprehensive assessment of the predictive performance of the machine learning models.

By accurately predicting reservoir porosity, the proposed framework assists in evaluating the suitability of geological formations for carbon dioxide storage. This supports effective reservoir characterization and improves the reliability of carbon capture and storage (CCS) assessment studies [3], [4], [11].

VI. RESULTS AND DISCUSSION

This section presents the experimental results and performance evaluation of the proposed machine learning framework for predicting reservoir porosity using well-log data. Several boosting ensemble algorithms were implemented and evaluated to determine the most effective model for accurate porosity prediction in geological formations. The evaluation focuses on comparing model performance, analysing prediction accuracy, and identifying the most influential well-log parameters contributing to porosity estimation.

A. Performance Comparison of Machine Learning Models

Several boosting ensemble machine learning algorithms were evaluated to determine the most suitable model for reservoir porosity prediction. The models implemented in this study include AdaBoost Regression, Gradient Boosting Regression, and Extreme Gradient Boosting (XGBoost). Model performance was evaluated using statistical evaluation metrics such as the coefficient of determination (R^2), Mean Squared Error (MSE), and Mean Absolute Error (MAE).

Table 1. Performance Comparison of Machine Learning Models

Model	**R^2 Score**	**MSE**	**MAE**
AdaBoost Regression	0.89	0.015	0.093
Gradient Boosting Regression	0.93	0.010	0.071

XGBoost Regression	0.96	0.007	0.054

From the comparison results, the XGBoost regression model achieved the highest prediction accuracy with an R^2 score of 0.96, outperforming the other ensemble algorithms. This improved performance can be attributed to the advanced gradient boosting mechanism of XGBoost, which efficiently captures complex nonlinear relationships between well-log parameters and reservoir porosity values. Additionally, the model effectively minimizes prediction errors through iterative optimization, resulting in improved predictive performance and model stability [5], [7].

B. Prediction Performance Analysis

The prediction capability of the proposed machine learning models was further analysed using statistical evaluation metrics that measure the difference between predicted and actual porosity values. The coefficient of determination (R^2) evaluates how well the model explains the variance in the dataset, while Mean Squared Error (MSE) and Mean Absolute Error (MAE) measure the prediction error.

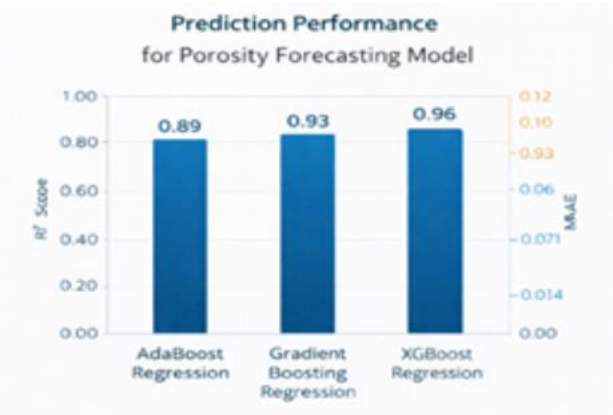

Fig. 2. Prediction Performance for Porosity Forecasting Model

The experimental results show that the boosting ensemble algorithms demonstrate strong predictive capability in estimating reservoir porosity from well-log data. Among the evaluated models, XGBoost achieved the lowest prediction error values and the highest R^2 score, indicating that the model provides more accurate and reliable predictions. The results demonstrate that boosting ensemble techniques are highly effective for modelling complex geological relationships present in subsurface reservoir datasets.

C. Feature Importance Analysis

To improve interpretability and understand the contribution of different well-log parameters, feature importance analysis was conducted to identify the most influential variables affecting porosity prediction.

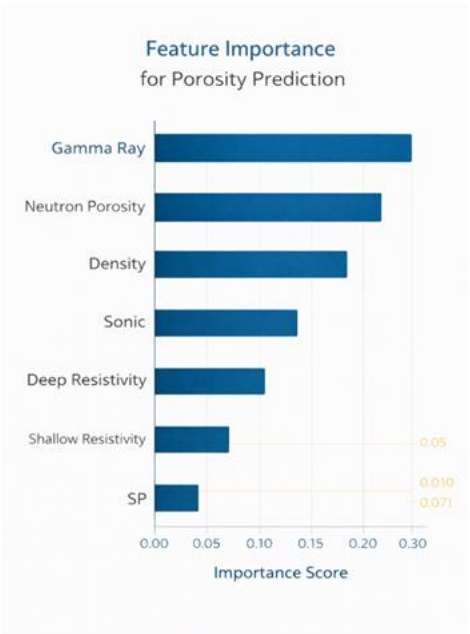

Fig. 3. Feature Importance for Porosity Prediction

The feature importance results indicate that several well-log parameters such as gamma ray, neutron porosity, density, and sonic logs significantly influence reservoir porosity estimation. Features with higher importance scores contribute more strongly to the prediction outcome, indicating their critical role in describing subsurface geological characteristics.

The feature importance analysis helps geoscientists and reservoir engineers understand which geological parameters most strongly affect porosity variations within the reservoir formation. This improves the transparency of the predictive model and enhances confidence in the model's predictions

for carbon dioxide storage assessment applications [1], [2], [8], [12].

VII. CONCLUSION AND FUTURE WORK

This study presented a machine learning framework for predicting reservoir porosity using well-log data to support carbon dioxide (CO_2) storage assessment. Accurate estimation of porosity is a critical factor in evaluating the storage capacity and suitability of geological formations for long-term carbon sequestration. However, traditional porosity estimation techniques such as laboratory core analysis are expensive, time-consuming, and limited to specific sample locations within the reservoir.

To address these challenges, a data-driven machine learning approach was developed using well-log measurements collected from subsurface geological formations. Several boosting ensemble algorithms, including AdaBoost Regression, Gradient Boosting Regression, and Extreme Gradient Boosting (XGBoost), were implemented and evaluated for predicting reservoir porosity. Among these models, the XGBoost regression model demonstrated the highest predictive performance, achieving superior accuracy with an R^2 score of approximately 0.96 while maintaining low prediction error values. This improved performance is attributed to the ability of boosting algorithms to capture complex nonlinear relationships between well-log parameters and reservoir properties [5], [7].

In addition, feature importance analysis was conducted to identify the most influential geological parameters affecting porosity prediction. The results revealed that well-log variables such as gamma ray, neutron porosity, density, and sonic logs play a significant role in determining reservoir porosity values. This analysis improves the interpretability of the predictive model and provides valuable insights for reservoir engineers and geoscientists involved in carbon storage evaluation [1], [2], [8].

The proposed framework provides an efficient and cost-effective alternative to traditional reservoir characterization techniques by utilizing machine learning algorithms and well-log datasets. By improving the accuracy and reliability of porosity prediction, the framework supports better decision-making in geological reservoir evaluation and carbon capture and storage (CCS) applications.

Future work may focus on integrating larger geological datasets from multiple reservoirs to improve model generalization and prediction robustness. In addition, advanced deep learning models and hybrid ensemble techniques may be explored to further enhance prediction accuracy. The integration of real-time drilling data and cloud-based reservoir monitoring systems could also enable automated reservoir characterization for large-scale carbon dioxide storage assessment.

REFERENCES

1. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
2. C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Munich, Germany: Leanpub, 2022.
3. S. Bachu, "CO_2 storage in geological media: Role, means, status and barriers to deployment," Progress in Energy and Combustion Science, vol. 34, no. 2, pp. 254–273, Apr. 2008.
4. IPCC, Special Report on Carbon Dioxide Capture and Storage. Cambridge, U.K.: Cambridge Univ. Press, 2005.
5. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794.
6. J. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.
7. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001.
8. G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning. New York, NY, USA: Springer, 2013.
9. A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2019.
10. R. J. Serra and M. J. Abrahamsen, "Machine learning techniques for reservoir property

prediction using well log data," Journal of Petroleum Science and Engineering, vol. 157, pp. 490–506, 2017.
11. J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks, vol. 61, pp. 85–117, Jan. 2015.
12. K. T. Uddin, M. Rahman, and S. M. Hossain, "Prediction of reservoir porosity using machine learning algorithms based on well log data," Energy Geoscience, vol. 3, no. 2, pp. 134–145, 2022.