

Research Article

Uncertainty-Driven Modeling of Microporosity and Permeability in Clastic Reservoirs Using Random Forest

Muhammad Risha¹, Mohamed Elsaadany², Paul Liu¹

1. Marine, Earth and Atmospheric Sciences Department, North Carolina State University, United States; 2. Geoscience Department, Universiti Teknologi Petronas, Malaysia

Predicting microporosity and permeability in clastic reservoirs is a challenge in reservoir quality assessment, especially in formations where direct measurements are difficult or expensive. These reservoir properties are fundamental in determining a reservoir's capacity for fluid storage and transmission, yet conventional methods for evaluating them, such as Mercury Injection Capillary Pressure (MICP) and Scanning Electron Microscopy (SEM), are resource-intensive. The aim of this study is to develop a cost-effective machine learning model to predict complex reservoir properties using readily available field data and basic laboratory analyses. A Random Forest classifier was employed, utilizing key geological parameters such as porosity, grain size distribution, and spectral gamma-ray (SGR) measurements. An uncertainty analysis was applied to account for natural variability, expanding the dataset, and enhancing the model's robustness. The model achieved a high level of accuracy in predicting microporosity (93%) and permeability levels (88%) using cross-validation and independent holdout evaluation. By using easily obtainable data, this model reduces the reliance on expensive laboratory methods, making it a valuable tool for early-stage exploration, especially in remote or offshore environments. By incorporating uncertainty, the model reflects geological variability more realistically and improves confidence in predictions, particularly when decisions must be made with limited or noisy input data. The integration of machine learning with uncertainty analysis provides a reliable and cost-effective approach for evaluating key reservoir properties in siliciclastic formations. This model offers a practical solution to improve reservoir quality assessments, enhancing decision-making in exploration planning and risk reduction, especially where direct measurements are limited enabling more informed decision-making and optimizing exploration efforts.

1. Introduction

Accurately predicting microporosity and permeability in clastic reservoirs can be crucial for assessing reservoir quality, particularly in complex geological environments where these properties are difficult and costly to measure directly^{[1][2]}. By utilizing machine learning models, these predictions can be made efficiently using readily available field data and inexpensive lab analyses, significantly improving decision-making in reservoir characterization. These parameters directly influence the capacity of a reservoir to store and transmit fluids, making them critical for hydrocarbon exploration and production^[3]. The presence of clay minerals such as kaolinite, illite, and chlorite further complicates this evaluation, as they can either preserve or reduce porosity and permeability through various diagenetic processes^[4]. Traditional laboratory methods such as Mercury Injection Capillary Pressure (MICP) and Scanning Electron Microscopy (SEM) provide detailed insights into pore structure and connectivity, but they are limited by high cost, destructive sample preparation, and limited field applicability. MICP requires specialized equipment and strict safety protocols, while SEM demands polished thin sections, vacuum environments, and extensive image interpretation. These factors restrict their use in large-scale reservoir studies or early-stage exploration settings, where rapid, non-destructive, and lower-cost alternatives are needed^[5]. Therefore, developing models that can predict these parameters under different geological conditions is a key focus of current reservoir studies^[6].

Several machine learning algorithms have been employed for predicting petrophysical properties, especially in the context of reservoir heterogeneity. In one study, lithology classification was conducted using a combination of self-organizing maps, neural networks, and genetic algorithms, which helped improve classification in mixed sand–shale successions^{[7][8]}. A separate approach utilized support vector machines to distinguish between marine and deltaic facies in seismic facies analysis using log-derived attributes^{[9][10]}. Random Forest classification was applied to effective reservoir identification using gamma-ray, resistivity, and acoustic logs, and demonstrated relatively stable performance in variable lithologies^{[7][11]}. Clastic facies classification was also conducted through ensemble ML models that incorporated gradient boosting, achieving moderate performance in distinguishing coarse- and fine-grained facies^[12]. A sediment-facies classification workflow combining feature selection with supervised learning improved the detection of muddy vs. sandy beds in mixed tidal flats^{[13][14]}. These studies

collectively point to the value of diverse algorithms when trained on appropriate input variables and tuned for geological settings.

Recent efforts have introduced more specialized ML strategies to address sparse labels, complex inputs, and prediction uncertainty. For instance, a semi-supervised method was used for tight sandstone gas prediction where only a fraction of the samples were labeled; the model leveraged consistency learning to maintain accuracy across unlabeled points^{[15][16]}. Another study integrated transfer learning with real seismic and synthetic facies data to improve classification performance in data-scarce depositional environments^{[12][17]}. In terms of model interpretability, explainable ML methods were applied to log-based prediction of shale porosity, using SHAP values to trace model decisions back to specific features^{[18][19]}. Deep learning was used to predict porosity from multiple seismic attributes, and the study emphasized probabilistic outputs to address uncertainty propagation^[20]. Review work has also identified trends across shoreline and reservoir ML workflows, showing how early empirical methods evolved into CNN-based spatial classification and LSTM-based time-series forecasting^[21]. These techniques highlight model adaptability when inputs are geophysically derived and uncertainty must be accounted for explicitly. Machine learning models have become powerful tools for addressing this challenge, providing the ability to process large datasets and capture complex relationships between geological, petrophysical, and diagenetic variables^[22]. By incorporating various inputs such as porosity, grain size distribution, and clay content, these models can predict the quality of reservoirs in clastic formations with a higher degree of accuracy^[23]. Specifically, Random Forest algorithms have proven effective due to their robustness in handling diverse data inputs and their ability to reduce overfitting, which is crucial when working with heterogeneous geological data^[24]. On top of that, Random Forest can handle missing values in predictor variables through imputation techniques, ensuring that incomplete datasets do not hinder predictive accuracy^[25]. This flexibility is vital in geological applications, where complete datasets are often difficult to obtain. The algorithm's strength in dealing with high-dimensional data and missing information makes it an ideal choice for predicting reservoir quality, even in data-limited scenarios^[26].

However, geological data inherently carries uncertainty due to variability in measurements and sample heterogeneity. For example, microporosity and permeability can vary significantly across different facies, complicating efforts to model these properties consistently^{[27][28]} (Figure 1).

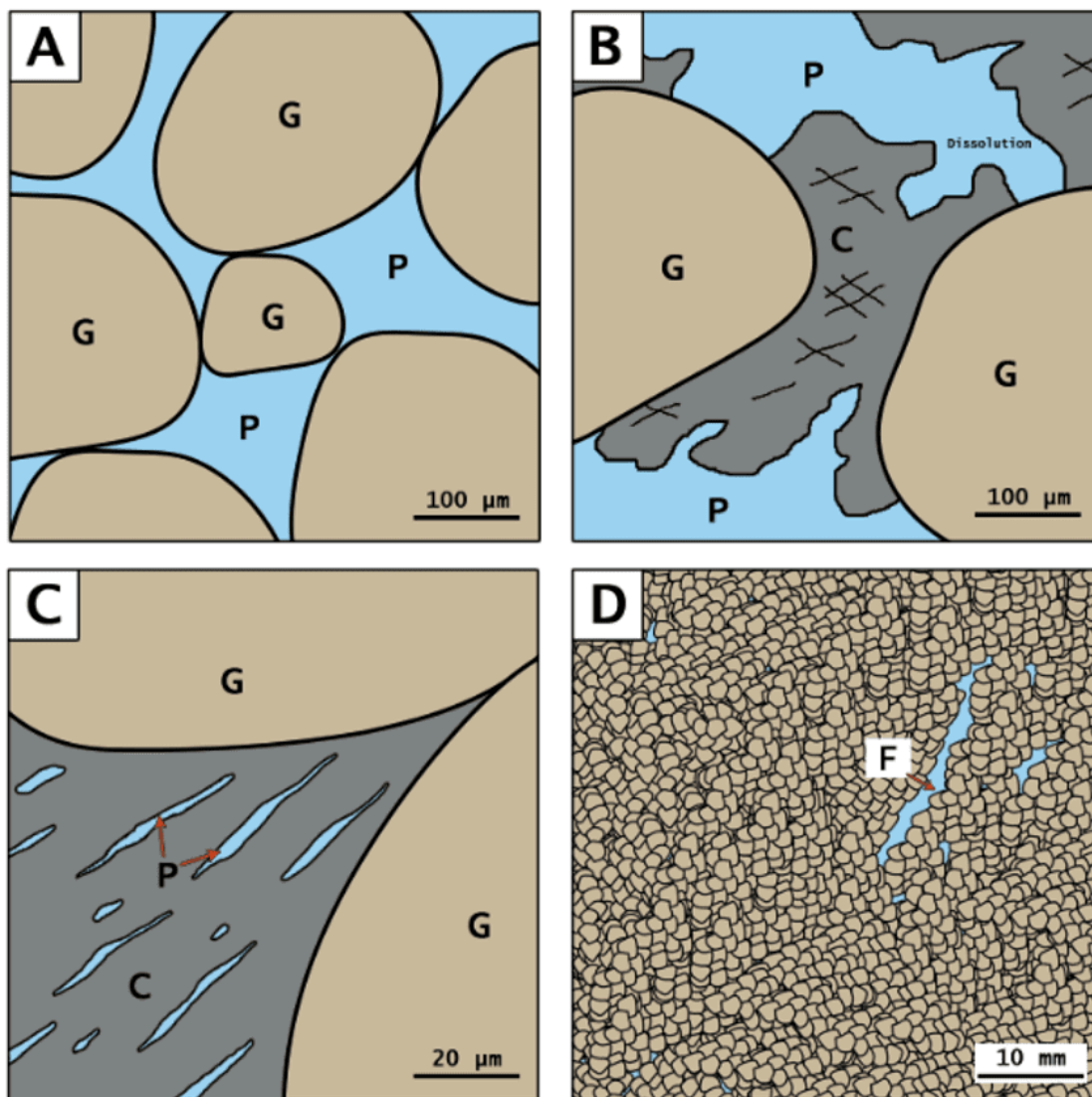


Figure 1. Schematic representation of the main porosity types in clastic reservoirs. (A) Intergranular porosity consists of primary pore spaces between detrital grains, providing essential storage and flow pathways. (B) Dissolution porosity forms when carbonate cement is partially dissolved, creating additional pore spaces that enhance reservoir quality. (C) Microporosity occurs within clay-rich matrices, where small pores retain fluids but significantly reduce permeability. (D) Fracture porosity develops when fractures cut through a compacted grain framework, improving permeability in otherwise low-porosity rocks. (G – Grains, P – Porosity, C – Clay, F – Fracture).

Addressing this variability is critical for creating models that not only predict the mean behavior of the reservoir but also account for the range of possible outcomes^[29]. To achieve this, uncertainty analysis can

be employed, allowing models to generate a broader set of predictions that better reflect the true range of geological conditions^{[30][31]}.

The research process involved field data collection, sample preparation, and laboratory analyses, followed by the integration of petrographic and petrophysical data to understand microporosity and permeability in clastic reservoirs. While all stages were crucial to the overall study, the focus of this paper is on the machine learning model development for predicting microporosity and permeability, based on the collected data (Figure 2). One of the key advantages of this model is its ability to predict difficult-to-obtain properties, such as microporosity and permeability, using basic field information and inexpensive laboratory analyses. Traditionally, these properties are measured through methods like MICP and SEM, which are both expensive and time-consuming. For example, laboratory-based permeability measurements often require core sampling and specialized equipment, which can take weeks or months to process and incur significant costs^[32]. Machine learning approaches offer a faster, more cost-effective alternative^[33] by leveraging widely available data such as porosity, grain size distribution, and gamma-ray logs.

While several ensemble and boosting algorithms such as XGBoost, CatBoost, and Gradient Boosting Machines (GBM) have shown high predictive accuracy in geoscientific and petrophysical applications, Random Forest was selected in this study due to its lower sensitivity to hyperparameter tuning and robustness when working with relatively small and heterogeneous datasets. RF does not require sequential tree building, unlike XGBoost and GBM, and is thus less computationally expensive and more straightforward to implement^[34]. A comparative evaluation across multiple classification tasks showed that although GBM offered slightly better accuracy, the difference in average performance between GBM, XGBoost, and RF was not statistically significant under default settings. RF also demonstrated stronger generalization with default parameters, making it suitable for cases where extensive hyperparameter optimization is impractical or constrained by computational cost^{[35][34]}. In addition, RF is more robust to overfitting and performs better on datasets with mixed data types or when dealing with high variance in features^{[36][37]}.

By integrating a robust uncertainty analysis, the model ensures that the variability in input data is adequately captured, further improving the reliability of predictions. This method not only reduces the need for expensive laboratory techniques but also enables the rapid prediction of reservoir properties from field data that can be collected more easily and affordably. Such an approach significantly reduces

the cost and time required for reservoir characterization, making it particularly valuable in exploration projects where access to advanced laboratory resources may be limited^[38].

The machine learning model developed in this paper has the potential to significantly reduce uncertainty in reservoir quality assessment, leading to better decision-making in hydrocarbon exploration.

2. Methodology

This study utilized field data collection, and laboratory analysis, used with machine learning techniques to predict microporosity and permeability in clastic reservoirs (Figure 2). The methodology involved gathering data from clastic outcrops on Labuan Island, Malaysia, which provided the key inputs for the machine learning model. The outcrops on Labuan Island, which form part of the larger Borneo geological structure, represent a range of formations, including the Crocker, Temburong, and Belait Formations, each characterized by unique depositional environments and diagenetic histories. These formations reflect the regional tectonic and sedimentological processes of the Sabah Basin, making the island an ideal location for collecting analog data applicable to offshore hydrocarbon exploration^{[39][40]}.

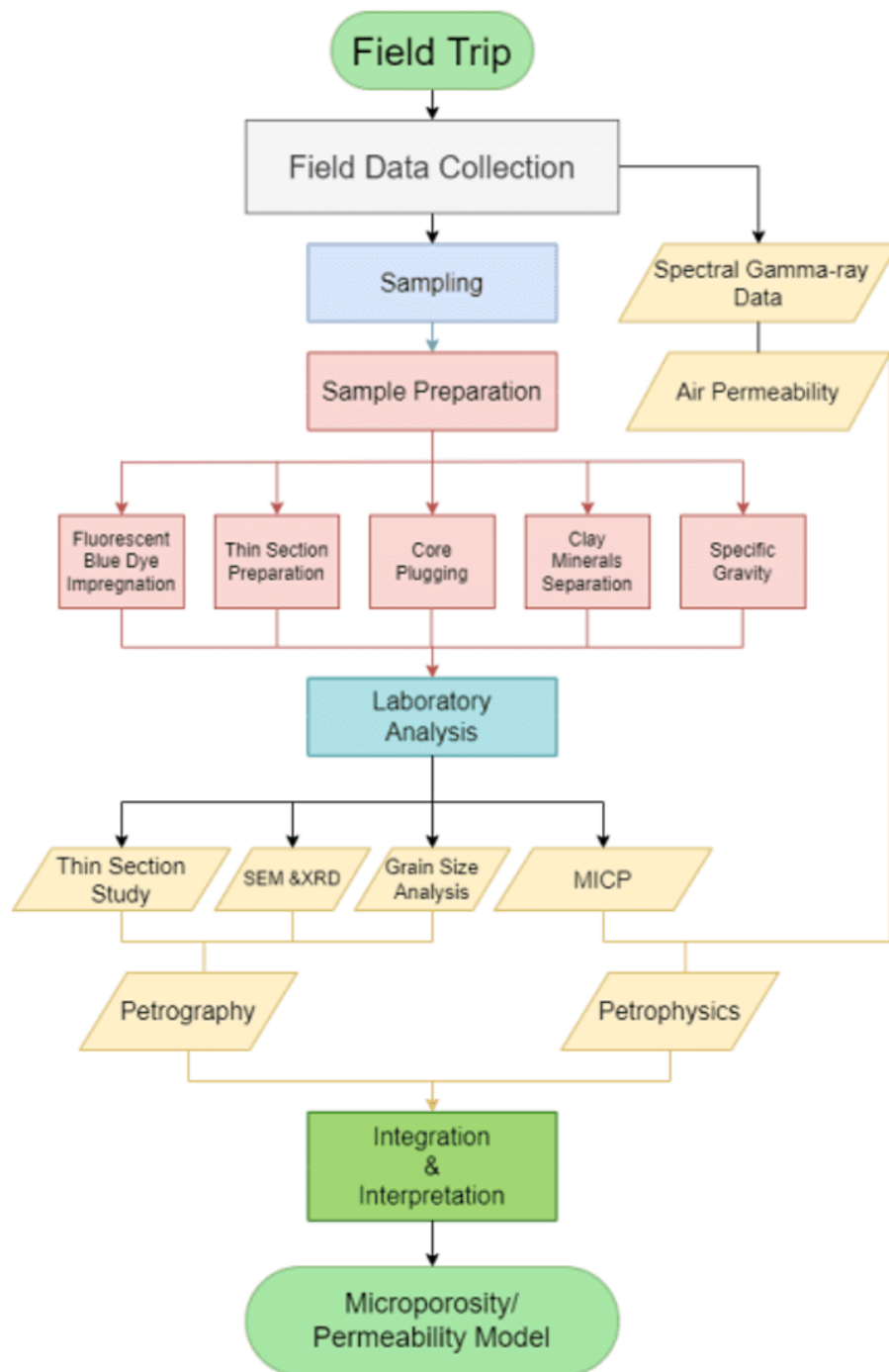


Figure 2. Workflow for data collection, preparation, and analysis leading to the development of the microporosity and permeability prediction model. The process includes field data collection, sample preparation, various laboratory analyses (e.g., MICP, SEM, XRD, and grain size analysis), and the integration of petrographic and petrophysical data to create a machine learning-based prediction model.

2.1. Data Collection and Preparation

Samples were collected from 16 outcrops across Labuan Island, Malaysia. The outcrops were selected based on accessibility, geographic and stratigraphic coverage, and lithological diversity, aiming to represent the main sandstone formations and facies on the island to ensure model generalization. Labuan is located off the northwest coast of Borneo and lies within the Brunei-Sabah Basin (Figure 3)^[41]. Its geology is dominated by clastic sequences deposited in deep to shallow marine environments, with varying clay content that influences reservoir quality through diagenetic alteration^[42]. The sampling targeted undisturbed exposures with minimal surface weathering. At each location, measurements were collected for total porosity from thin sections, air permeability, and grain size distribution. To improve reproducibility, coordinates were recorded for all outcrops and are shown in Figure 3, allowing future studies to locate and revisit the sampling sites under similar conditions.

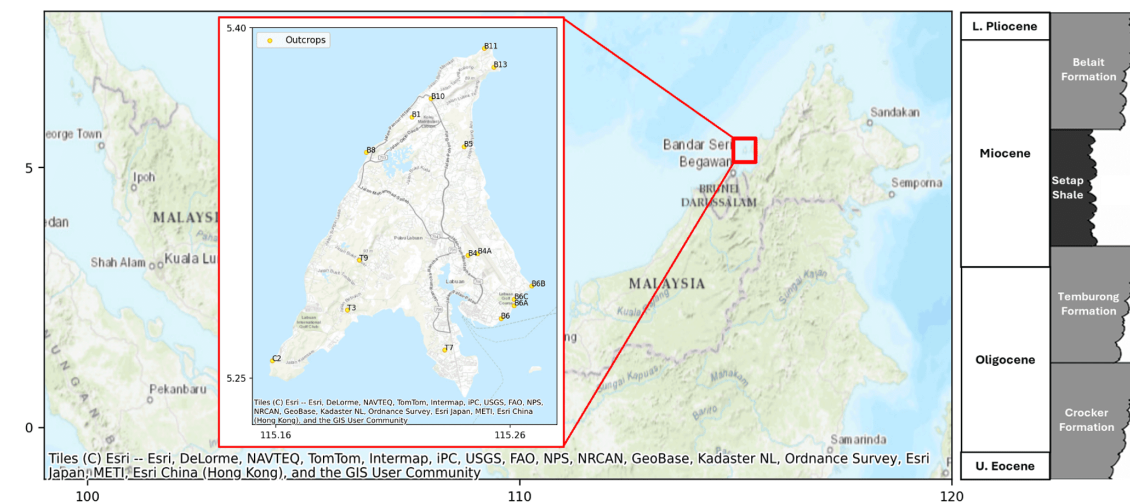


Figure 3. Regional and geological context of the study area. The maps show study area, and selected outcrop locations across Labuan Island, highlighted within a regional context. The stratigraphic column illustrates the relative chronostratigraphic succession of geological formations on the island, ranging from the Upper Eocene to the Late Pliocene, providing insights into its depositional history^{[43][44]}.

Porosity was measured using multiple techniques to ensure comprehensive analysis. Thin section analysis was conducted on rock samples to visually estimate porosity under a polarized microscope using blue impregnated epoxy which is helpful in identifying pore spaces. Additionally, Helium porosimetry was employed as well to measure porosity, providing a more accurate estimation of the interconnected

pore spaces within the rock samples^[45]. MICP analysis was another key method used to quantify microporosity^[46]. It has provided critical insights into the pore sizes distribution, particularly in identifying pore sizes smaller than 2 μm in this research.

Permeability was measured using two approaches. First, air permeability was assessed in both horizontal and vertical orientations using the TinyPerm II Air-permeameter. This field-based method provided a quick, cost-effective and non-destructive estimation for permeability directly at the outcrop. Horizontal and vertical measurements were taken to account for anisotropy in permeability due to layering and sedimentary structures within the rock. Second, laboratory-based MICP analysis was conducted to obtain more detailed permeability data, particularly in tight rock samples where microporosity dominates^[29].

Clay, sand, and silt content were quantified using a combination of sieve analysis for the coarser fractions and hydrometer analysis for the finer particles as an assessment of the grain size distribution within each sample. SGR analysis was also performed on the outcrops to provide estimates of the clay content in terms of type, with specific attention to clay minerals like kaolinite, illite, and chlorite, which are known to influence porosity and permeability^[6]. XRD and SEM were also used for few samples to identify the specific types of clay minerals and their spatial distribution within the pore spaces^[47].

The data collected from these various field and laboratory methods formed the foundation for developing the machine learning model. Each measurement including porosity, permeability, or clay content was carefully chosen to reflect the primary factors influencing reservoir quality in clastic rocks. The integration of these diverse datasets ensures a comprehensive understanding of the factors controlling microporosity and permeability in the formations studied.

2.2. Uncertainty Analysis

To enhance the machine learning model's robustness and its ability to generalize similar geological conditions, an uncertainty analysis was applied. Model Generalization allows the model to perform better on unseen datasets. This process allowed for a broader representation of feature variability, ensuring the model could better capture the natural heterogeneity of clastic formations^[48].

This uncertainty analysis played a crucial role in preparing the machine learning model to handle real-world variability. By training the model on a broader and more representative dataset, the analysis ensured that the model was better equipped to predict microporosity and permeability under similar range of geological conditions without the necessity of data from the same outcrops. Incorporating these

uncertainty ranges minimized the risk of overfitting, enabling the model to produce more reliable and accurate predictions for reservoir quality^{[49][50]}.

Key input parameters subject to uncertainty included total porosity, air permeability, sand, silt, and clay content, the thorium-potassium (ThK) ratio, the sand-to-mud (SM) ratio, mean grain size, standard deviation, skewness, permeability, and microporosity ($<2\mu\text{m}$). These parameters were chosen due to their significant influence on reservoir quality and their role in controlling fluid flow and storage capacity.

Uncertainty-based resampling was used to simulate natural geological variability and analytical uncertainty in the dataset. For each original sample, additional synthetic datasets were created by perturbing key input parameters within predefined uncertainty ranges (Table 1). These perturbations followed a uniform random distribution and reflected average measurement variability, human estimation error, and instrument precision. Derived variables such as grain size statistics and SM ratio were recalculated from the adjusted values to maintain internal consistency. This approach is widely recognized in geological and environmental modeling and is conceptually related to bootstrap resampling and Monte Carlo simulations^{[51][52]}, both of which introduce random variation to explore the uncertainty space around observed data.

The resampling process generated 200 synthetic datasets per sample, applied across 308 original records from 41 samples, resulting in a final training set of 61,600 data points. This expanded dataset allowed the Random Forest model to learn from a broader range of plausible geological conditions, improving its generalization capacity and robustness. The integration of uncertainty into machine learning workflows is increasingly supported in recent literature as a way to reduce overfitting and improve model reliability under variable input conditions^[53].

Parameter	Uncertainty Range	Units	Remarks
Porosity	±4%	Percentage (%)	Based on average variability.
Air Permeability	±10	Millidarcies (mD)	Based on average variability.
Clay Content	±2%	Percentage (%)	Adjusted to simulate error of sieve and hydrometer tests. Their total has to be 100%.
Sand Content	±2%	Percentage (%)	
Silt Content	±2%	Percentage (%)	
ThK Ratio	±0.5	Ratio	Based on average variability.
SM Ratio	Recalculated	N/A	Recalculated from updated values.
Mean Grain Size	Recalculated	Millimeters (mm)	Based on human estimation error.
Standard Deviation	Recalculated	N/A	Based on human estimation error.
Skewness	Recalculated	N/A	Based on human estimation error.
Target Permeability	±10	Millidarcies (mD)	Based on average variability.
Target Microporosity	±5%	Percentage (%)	Based on average variability.

Table 1. Predefined Uncertainty Ranges in the Uncertainty resampling Model.

2.3. Machine Learning Model Development

This study employed a Random Forest classifier to predict microporosity and permeability levels in clastic reservoirs. Random Forest was selected because of its ability to handle complex, multi-dimensional geological data with both continuous and categorical variables. By aggregating the results of multiple decision trees, Random Forest reduces the likelihood of overfitting and enhances prediction accuracy, especially when dealing with noisy geological data^[54].

To evaluate the influence of each input variable on the predictive performance of the classification models for permeability and microporosity, a feature importance analysis was conducted using the Gini importance metric. This metric assesses the total decrease in node impurity (Gini impurity) attributed to each feature across all trees in the ensemble, providing a relative measure of each variable's contribution to the model's decision-making process^[55]. In this analysis, the computed importance scores were normalized and visualized using heatmaps, where features were ranked in descending order of importance. Color intensity in the heatmaps reflects the relative importance of each input variable, with darker tones indicating stronger influence (Figure 4). This approach facilitates the identification of key predictors and enhances the interpretability of the model's internal mechanisms, aligning with established practices in machine learning applications within geosciences^{[56][57]}.

The model utilized input variables that are crucial to reservoir quality assessment, including porosity, air permeability, clay content, grain size distribution, ThK ratio, sand-to-mud (SM) ratio, and the graphic sedimentary parameters. The model's target variables were permeability categories and microporosity domination (Figure 5). Permeability was classified into three categories: (Poor to Fair), (Moderate), and (Good to Very Good) (Table 2) (Adaeze et al., 2012). This classification helps in distinguishing between different production potentials of the reservoir. Additionally, microporosity domination was used to classify samples as either macroporosity or microporosity-dominated, a critical distinction that affects the fluid flow within the reservoir. Porosity and permeability were key variables as they directly influence the reservoir's ability to store and transmit fluids. The ThK ratio is a proxy for clay content, and the grain size distribution, which includes sand, silt, and clay for the rock's texture.

Permeability Range	Qualitative Description
Poor to Fair	<15
Moderate	16-50
Good to V. Good	<50

Table 2. Qualitative evaluation for permeability ranges

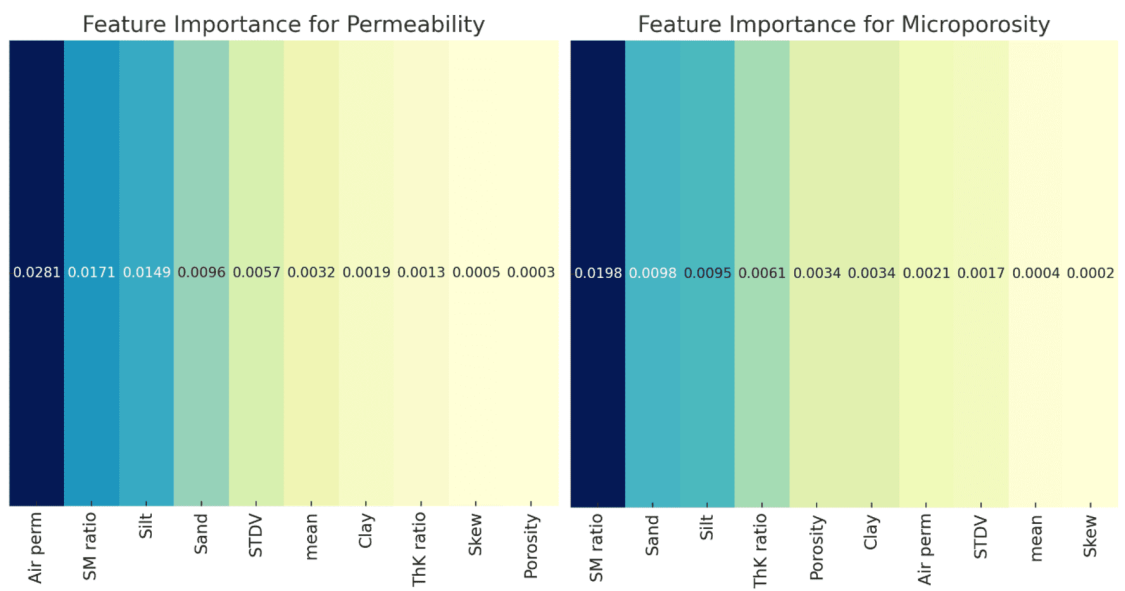


Figure 4. Ranked feature importance for permeability (left) and microporosity (right) models. Color intensity reflects the relative importance of each input variable, with darker tones indicating stronger influence. Only physical and compositional variables were included.

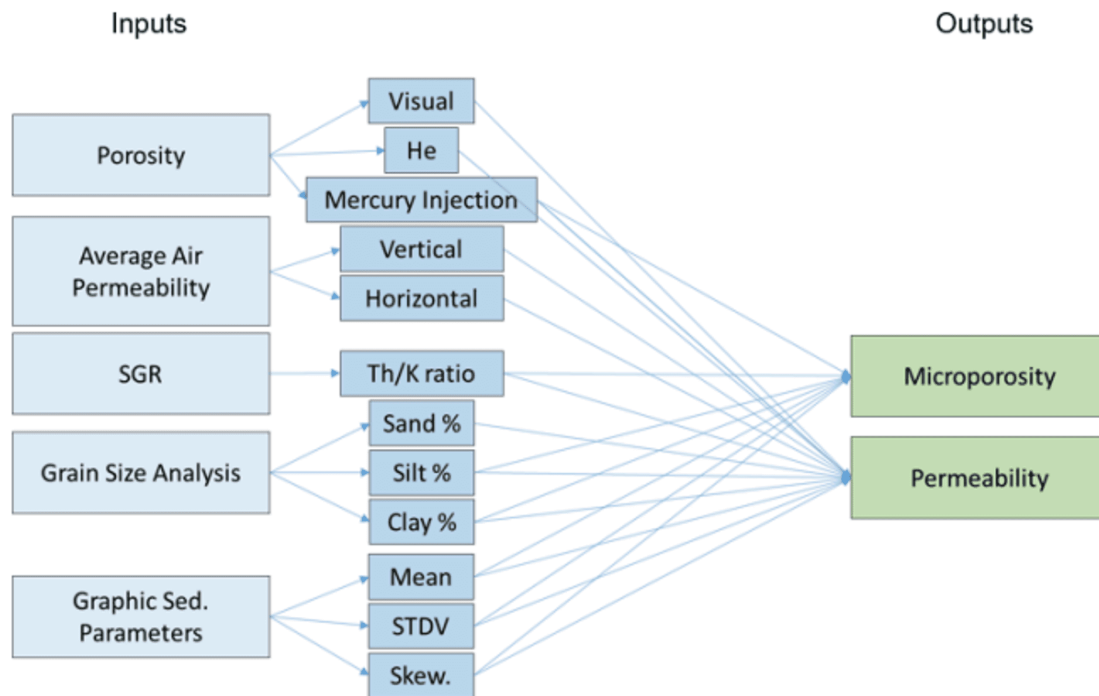


Figure 5. A schematic representation of the Random Forest model inputs and outputs. The inputs consist of Porosity, Air Permeability, Th/K Ratio, Grain size data, and Graphic sedimentary parameters, while the outputs are descriptive permeability level and microporosity domination ($<2\mu\text{m}$).

The Random Forest model was configured with 100 decision trees, each limited to a maximum of five leaves. A decision tree works by splitting the dataset into subsets based on feature values, which can be thought of as a sequence of "if-then" rules. Each branch of the tree represents a decision, and the leaves are the final predictions. By limiting the maximum number of leaves to five, it was ensured that the model can avoid overfitting. With a large number of shallow trees (100 trees with 5 leaves each), the model captures the key relationships between input variables while avoiding learning noise or irrelevant patterns^{[58][59]}. The model's hyperparameters, including the number of trees, maximum depth, and minimum samples per leaf, were selected through grid search optimization. Multiple configurations were evaluated using cross-validation, and the final parameters were chosen based on the lowest validation error and reduced overfitting. This approach ensured that the model remained generalizable while maintaining high predictive performance.

The final training process used 90% of the dataset for training the Random Forest model, while the remaining 10% was reserved for evaluation. Training refers to the phase in which the model learns from

the data by identifying patterns between the input features (such as porosity, permeability, and clay content) and the target variables (permeability and microporosity categories). By exposing the model to this portion of the data, it can "learn" relationships and develop the ability to make predictions on unseen data. Another testing phase was done by applying the model on a completely separate holdout dataset to evaluate how well the trained model performs on unseen data to check the model's ability to generalize to real-world cases.

To validate the model during the training phase and prevent overfitting, 5-fold cross-validation was applied. In this technique, the training data (90%) was divided into five equal parts, or (folds). The model was then trained on four of these folds and validated on the remaining fold. This process was repeated five times, with each fold serving as the validation set once. The average performance across all five iterations was recorded. Cross-validation not only enhances the model's reliability but also reduces the risk of overfitting, as each portion of the data is used for both training and validation at different stages^{[60][61]}. The 5 cross-validation was applied to ensure that performance was evaluated across different subsets of the data, reducing the influence of individual points. In addition, uncertainty-based resampling expanded the training set with realistic variability, which further diluted the impact of extreme values and improved model robustness. The Random Forest algorithm was selected in part due to its ability to maintain performance in the presence of variable input data^{[62][63][55]}.

Before training the model, categorical variables, such as permeability and microporosity levels, were converted into numeric representations. This process, called encoding, was necessary because the machine learning algorithm operates on numerical data. Each categorical level was assigned a numerical code, allowing the Random Forest algorithm to process and learn from these features effectively. The encoding ensured that the model could properly interpret permeability and microporosity levels, which were crucial target variables for prediction^[64].

To assess the performance of the Random Forest classifier, the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) metrics were employed. The ROC curve provides a graphical representation of the model's diagnostic ability by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. Mathematically, the TPR (also known as sensitivity or recall) is calculated as the ratio of correctly predicted positive observations to the total actual positives, while the FPR is the ratio of incorrectly predicted positive observations to the total actual negatives^[65]. The AUC value, which ranges from 0 to 1, serves as a scalar measure summarizing the model's overall discriminatory ability, with values closer to 1 indicating the highest performance. In this study, the ROC

curve was calculated for each class within the target variables using the predicted probabilities produced by the Random Forest model^{[66][67]}. These probabilities, rather than binary classifications, allowed for a more nuanced evaluation of the model's performance across a spectrum of thresholds.

After completing the training and cross-validation steps, the holdout set was evaluated using the model to ensure an unbiased evaluation of the model's generalization capabilities.

3. Results

This section presents the results of the Random Forest model's performance in predicting permeability and microporosity domination, evaluated through cross-validation, an internal 10% evaluation set, and an independent holdout dataset. The model's accuracy was measured using confusion matrices, and the impact of uncertainty in the input parameters was also analyzed.

3.1. Model Accuracy

The model was trained to classify both permeability levels and microporosity domination using Random Forest. Classification performance was evaluated on a 10% subset excluded from training. Without uncertainty resampling, the accuracy was 86.4% for permeability and 88.2% for microporosity. After incorporating uncertainty, the corresponding values increased to 87.8% and 92.7%, respectively. Class-wise prediction results are shown in Figure 6. For permeability, post-uncertainty classification accuracy reached 92.8% for "Poor to Fair," 72.5% for "Moderate," and 98.0% for "Good to Very Good." For microporosity, accuracy was 90.3% for macroporosity-dominated samples and 94.1% for microporosity-dominated samples.

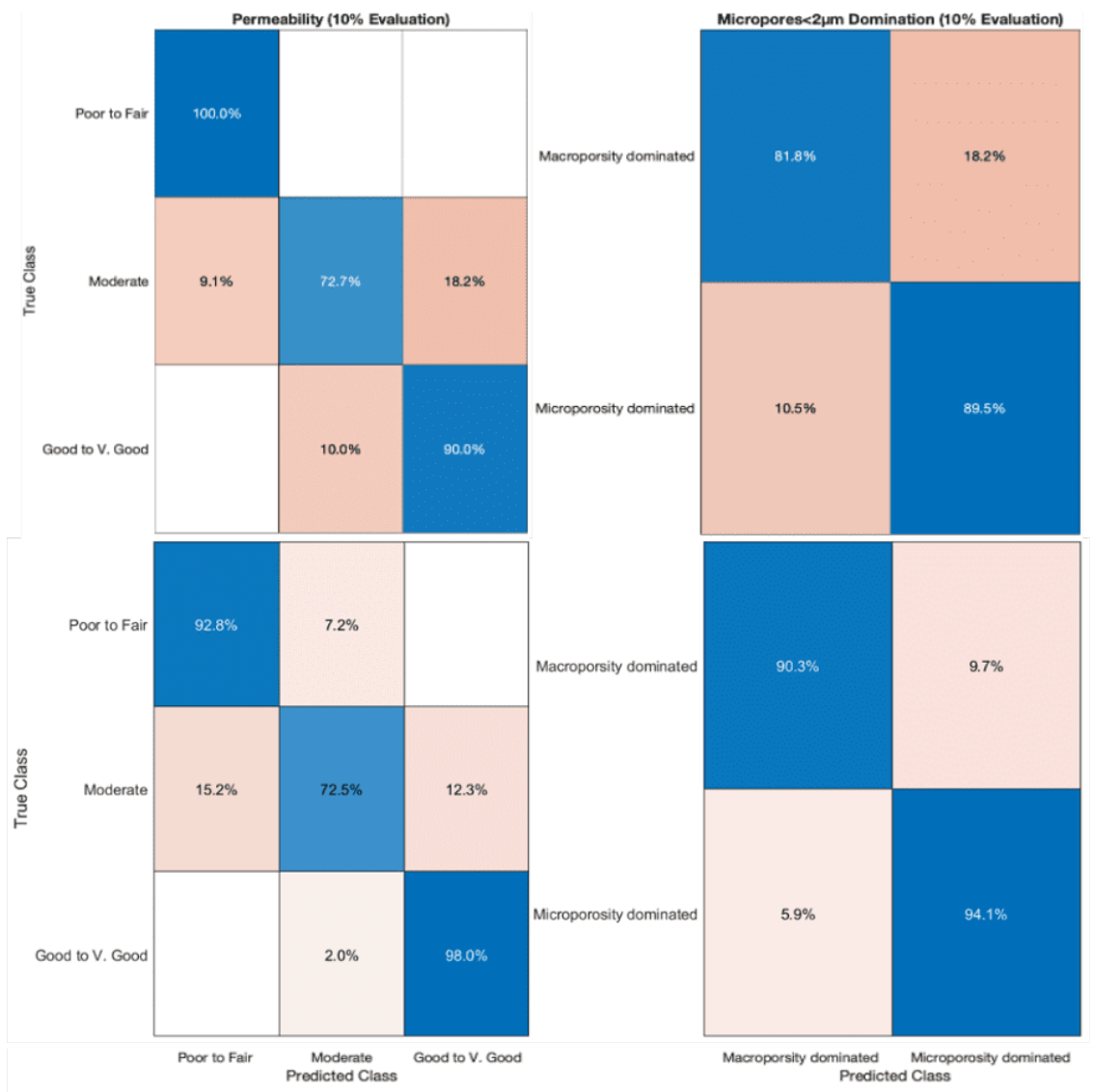


Figure 6. Confusion matrices showing model predictions before (top row) and after (bottom row) applying uncertainty-based resampling. (Left) Permeability classification results. (Right) Microporosity domination classification results. Each matrix reports the percentage of correctly and incorrectly classified samples for each class based on a 10% evaluation set.

3.2. Uncertainty Impact

The effect of uncertainty incorporation was assessed using both cross-validation error and evaluation accuracy. Without uncertainty resampling, the cross-validation error was 19.51% for permeability and 9.87% for microporosity. These values were reduced to 11.83% and 6.76%, respectively, after uncertainty was applied. To further illustrate the effect, class-wise accuracy before and after uncertainty

incorporation is presented in Figure 7. Uncertainty resampling introduced variability within observed ranges for each parameter, increasing the number of training instances and stabilizing the classifier's response. This helped reduce overfitting and improved generalization to new data.

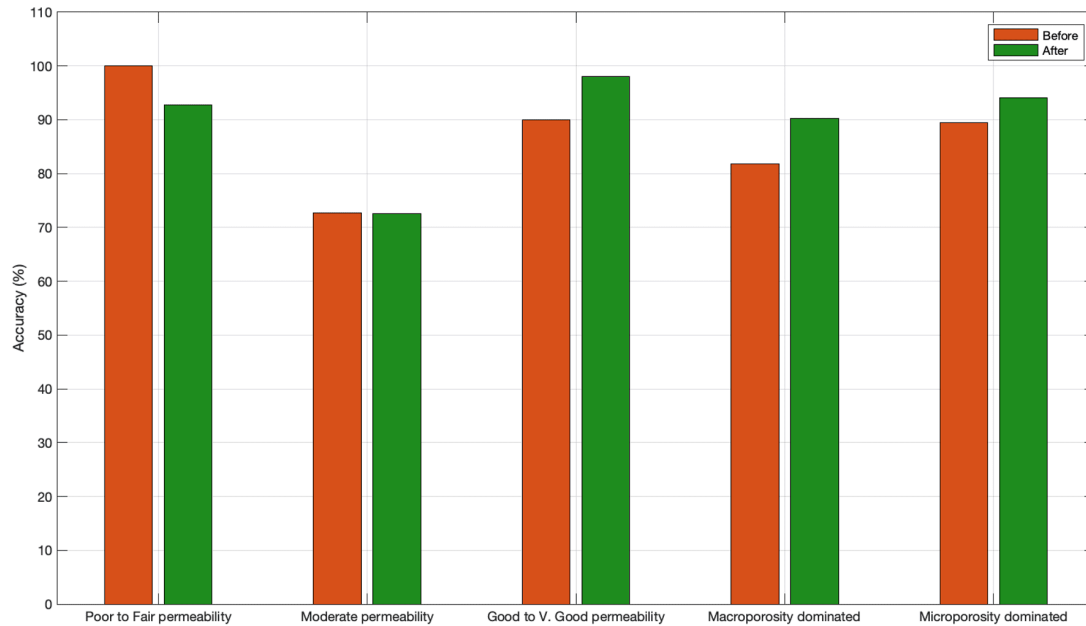


Figure 7. Bar chart as a performance comparison as classification accuracy before and after applying uncertainty-based resampling. Bars include five classes: "Poor to Fair permeability," "Moderate permeability," "Good to Very Good permeability," "Macroporosity dominated," and "Microporosity dominated." Accuracy values are based on the 10% evaluation set.

3.3. Holdout Set Evaluation

The independent holdout dataset consisted of 11 samples that were not used in either model training or cross-validation. These samples were selected to represent a range of lithological and petrophysical conditions and serve as a final check for model generalization. The classification accuracy reported below is based solely on these unseen samples.

The model's generalization capability was further tested using an independent holdout dataset that was not used during training or internal evaluation. The evaluation accuracies on this independent holdout dataset were 90.91% for permeability and 81.82% for microporosity domination.

Permeability type showed perfect 100% accuracy predicting (Poor to Fair) and (Moderate) permeability, while misclassifying only one sample of (Good to V. Good) as (moderate) permeability calculated as 80% accuracy. For porosity type domination, the confusion matrix shows that the model correctly predicted 90.3% of the macroporosity-dominated samples, with 9.7% misclassified as microporosity-dominated. For the microporosity-dominated samples, 94.1% were correctly predicted, with 5.9% incorrectly predicted as macroporosity-dominated (Figure 8).

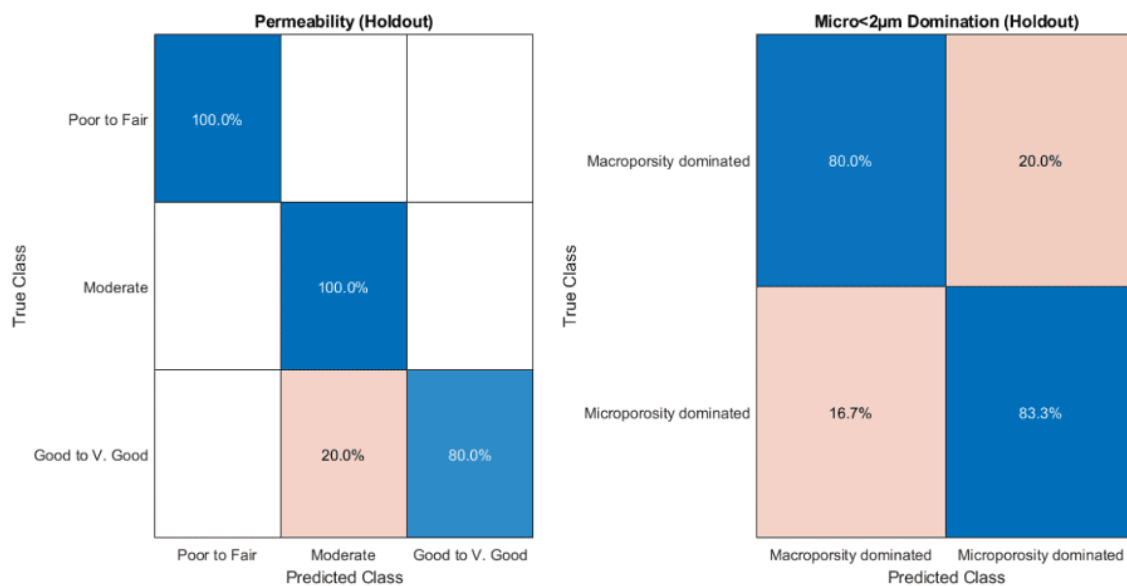


Figure 8. Confusion matrices for permeability and microporosity predictions using the Random Forest model (Holdout Set). (Left) The confusion matrix for permeability shows 100% accuracy in predicting both "Poor to Fair" and "Moderate" permeability categories, with slight misclassification in the "Good to Very Good" category (80% accuracy). The model performed very well on the holdout set, demonstrating strong generalization. (Right) The confusion matrix for microporosity predictions displays good accuracy, with 83.3% of "Microporosity dominated" cases correctly classified, though there was some misclassification between "Macroporosity dominated" and "Microporosity dominated" classes (80% and 16.7% misclassification, respectively). The model's performance on the holdout set reflects its capability to generalize predictions to new, unseen data.

4. Discussion

This section provides an interpretation of the model's results, focusing on the importance of the uncertainty analysis, the implications of the model's accuracy, and potential areas for improvement in future work.

4.1. Model Accuracy and Implications

The high accuracy of the model in predicting permeability and microporosity has significant implications for reservoir quality assessment. With cross-validation errors of 11.83% for permeability and 6.76% for microporosity, the model demonstrated strong predictive power. Evaluation on the holdout dataset resulted in accuracies of 90.91% for permeability and 81.82% for microporosity domination, confirming the model's generalization capabilities.

The model demonstrates strong predictive performance in estimating permeability and microporosity categories, as reflected by the high AUC values. For permeability prediction, AUC scores of 0.97, 0.90, and 0.98 for the (Poor to Fair), (Moderate), and (Good to Very Good) categories, respectively, highlight the model's ability to distinguish between different reservoir quality levels with high accuracy. Similarly, microporosity classification achieves an AUC of 0.95 for both macroporosity- and microporosity-dominated samples, reinforcing the reliability of the model in differentiating pore types (Figure 9). The consistently high AUC values suggest robust classification performance across varying thresholds, making the model a valuable tool for reservoir characterization and predictive analysis.

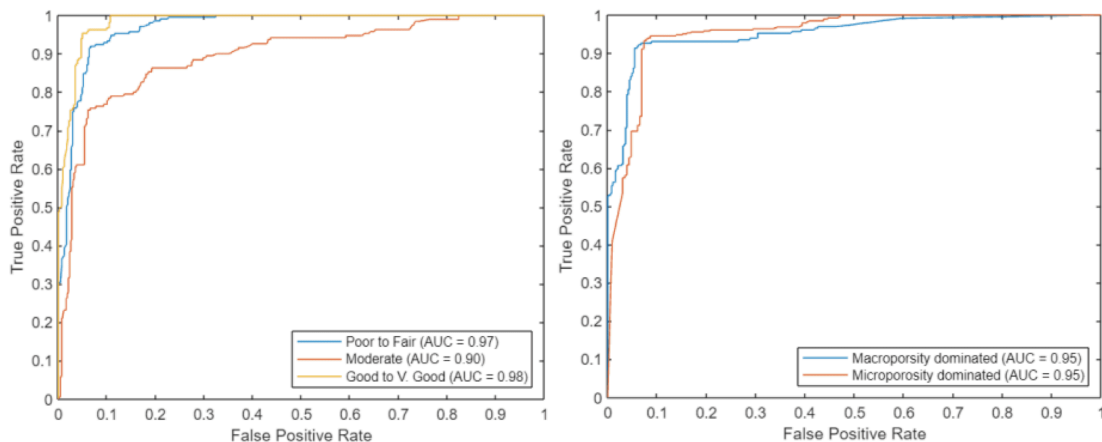


Figure 9. ROC curves illustrating the performance of the Random Forest classifier for predicting permeability categories and microporosity domination. The curves depict the trade-off between the true positive rate and false positive rate at various threshold levels. AUC values, displayed in the legend, provide an overall measure of the model's accuracy in distinguishing between different classes.

Table 3 shows how well the model predicted both permeability and microporosity domination for an independent holdout dataset. Most of the predictions align with the actual values, indicating the model's

strong performance. For instance, samples C2-4 and C2-6 were correctly identified as (Poor to Fair) in terms of permeability, while samples like T3-2 and B6C-4 were accurately classified as (Good to V. Good).

When it comes to microporosity domination, the model also did well, correctly predicting samples such as C2-4, C2-6, and T9-1 as (Microporosity dominated). However, there were a few cases of misclassification. For example, sample B13-1 was predicted as (Good to V. Good) when it was actually (Moderate), and sample B6C-7, which was (Microporosity dominated), was incorrectly predicted as (Macroporosity dominated).

These misclassifications point out some of the model's limitations. It seems to struggle with certain subtleties in distinguishing between the classes, possibly due to complex overlaps in the dataset's features. Despite this, the majority of predictions were spot-on, showing that the model can be a reliable tool for assessing reservoir quality. It offers a solid foundation for future improvements, such as adding more detailed features or refining the model to better capture the nuances of microporosity and permeability.

Sample	Actual Permeability	Predicted Permeability	Actual Porosity	Predicted Porosity
C2-4	Poor to Fair	Poor to Fair	Microporosity dominated	Microporosity dominated
C2-6	Poor to Fair	Poor to Fair	Microporosity dominated	Microporosity dominated
T3-2	Good to V. Good	Good to V. Good	Macroporosity dominated	Macroporosity dominated
B6C-3	Good to V. Good	Good to V. Good	Macroporosity dominated	Macroporosity dominated
B6C-4	Good to V. Good	Good to V. Good	Macroporosity dominated	Macroporosity dominated
B6C-7	Good to V. Good	Good to V. Good	Microporosity dominated	Macroporosity dominated
T8-2	Poor to Fair	Poor to Fair	Microporosity dominated	Microporosity dominated
T9-1	Poor to Fair	Poor to Fair	Microporosity dominated	Microporosity dominated
B10-2	Moderate	Moderate	Microporosity dominated	Microporosity dominated
B11-3	Moderate	Moderate	Microporosity dominated	Microporosity dominated
B13-1	Moderate	Good to V. Good	Macroporosity dominated	Macroporosity dominated

Table 3. MICP Actual vs Predicted Values for the independent holdout dataset. Green cells show the correct predictions while red cells show wrong predictions.

The uncertainty ranges played an important role in boosting the model's accuracy by incorporating the natural variability found in geological data. By training the model on a dataset that included this variability, it was exposed to a broader range of conditions, which helped to reduce the impact of measurement inconsistencies and prevent overfitting. This approach enabled the model to better adapt to new, unseen data. The inclusion of uncertainty analysis resulted in a significant reduction in prediction error, with approximately a 40% improvement for permeability and about a 30% improvement for microporosity. These findings underscore the importance of considering data variability in modeling, leading to more robust and reliable predictions, particularly in complex geological contexts. This demonstrates that incorporating such analysis is not merely an optional step but a crucial component in refining model performance for practical applications.

When compared to both simpler and more complex machine learning methods, the Random Forest model demonstrated superior performance in this study. Its ensemble learning approach provides greater stability and accuracy in predictions, which simpler models like linear regression or decision trees may struggle to achieve, particularly in the presence of non-linear geological data^[68]. While more complex models like neural networks can sometimes offer higher accuracy, they often require significantly larger datasets and more computational resources. Random Forest strikes an effective balance, reducing the risk of overfitting while still handling heterogeneous datasets and missing data with robustness^[26].

4.3. Limitations and Future Improvements

Random Forest is generally considered robust against overfitting compared to many other machine learning algorithms, including decision trees and deep learning models. Its use of ensemble learning, bootstrapped sampling, and randomized feature selection contributes to low model variance and stable predictions even in complex or noisy datasets^{[69][63]}. Studies have shown that Random Forest maintains consistent performance across a range of hyperparameter settings and is particularly resilient when working with high-dimensional geological data or limited sample sizes^{[62][70]}. While overfitting is not a major concern, the model can still underperform on rare or poorly represented geological cases. This is not due to memorization of the training data, but rather the lack of sufficient patterns to generalize from. Addressing this limitation will require a broader dataset that captures more edge-case lithologies and potentially hybrid models that better handle minority-class behavior.

While the Random Forest model demonstrated strong performance, there are several limitations that should be addressed in future work. One key limitation is the model's sensitivity to the quality of input data. For example, misclassifications of microporosity domination in samples B6C-7 and T8-2 were likely due to the model's difficulty in handling laminated and interbedded sandstone facies. These facies were not directly included in the model as numerical features, which led to misclassifications. Incorporating more detailed sedimentological information into the model could improve its accuracy in these cases.

Additionally, the "Moderate" permeability class showed lower classification accuracy compared to the other categories. This may be due to overlap in input features between adjacent classes, especially when geological characteristics are transitional or not sharply defined. In these cases, the boundaries between "Poor to Fair," "Moderate," and "Good to Very Good" are not distinct, making the classification task more difficult for the model. Increasing the number of training samples in this range and incorporating more

detailed facies descriptors or transitioning to regression-based approaches could help reduce this uncertainty.

Sedimentological properties such as facies type have a significant impact on reservoir characteristics, but they are difficult to quantify numerically. Developing methods to integrate facies data into machine learning models, either through the use of numerical facies codes or by incorporating more detailed geological parameters, could enhance model accuracy and reduce misclassifications.

Future work could also explore the use of other machine learning algorithms, such as gradient boosting machines or neural networks, which may offer improvements in handling more complex relationships between variables. Additionally, increasing the diversity of the training dataset by including more geological formations and a wider range of sedimentological properties could further enhance the model's generalization capabilities.

The integration of uncertainty analysis with advanced machine learning techniques, combined with more detailed geological data, could lead to the development of even more accurate models for predicting reservoir quality. This would provide a valuable tool for both academic research and practical applications in the oil and gas industry.

5. Conclusion

The Random Forest machine learning model developed in this study successfully predicted permeability and microporosity domination ($<2\ \mu\text{m}$) with high accuracy, achieving 90.91% for permeability and 81.82% for microporosity. By using key geological parameters such as total porosity, SGR data, and grain size distribution, the model proved to be a reliable tool for predicting reservoir quality in siliciclastic formations.

A major benefit of the model is its ability to predict complex properties like microporosity and permeability, which typically require expensive, time-consuming methods. Instead, the model leverages basic, cost-effective data such as porosity and grain size analysis, making it ideal for early-stage exploration and reducing both cost and uncertainty.

Incorporating uncertainty analysis enhanced the model's robustness, allowing it to generalize well across different formations. This approach is particularly useful in remote or offshore settings, where access to advanced laboratory testing is limited. The ability to generate accurate predictions using affordable data improves decision-making in exploration projects.

For future improvements, incorporating geological parameters like facies could increase prediction accuracy, and advanced imaging techniques such as backscattered SEM or microCT may provide more detailed insights into microporosity. Exploring different microporosity thresholds and expanding the dataset with more well logs could further enhance the model's performance and generalizability across different formations. Additional extensions could also include testing other machine learning algorithms such as XGBoost and Support Vector Machines to evaluate their effectiveness in capturing nonlinear patterns. Expanding the diversity of the input data, particularly with respect to lithology and reservoir heterogeneity, may also improve the model's applicability to broader depositional settings.

In conclusion, the Random Forest model, combined with uncertainty analysis, provides a cost-effective and powerful framework for predicting permeability and microporosity in clastic formations. By applying this model to similar geological contexts, reservoir quality predictions can be significantly improved, reducing uncertainty and optimizing exploration efforts, particularly when resources for expensive laboratory analyses are limited.

Statements and Declarations

Author Contributions

Conceptualization: MR; Literature Review: MR; Methodology: Lab Analysis MR; MR; Software Development and Validation: MR; Writing – original draft: MR; Writing – review & editing: MR, ME, and PL. The authors have read and approved the final manuscript.

Funding

This research was funded by the YUTP grant (015LC0-060).

Data Availability

The data supporting the findings of this study are available on reasonable request from the corresponding author.

Conflicts of interest

The author declares no relevant financial or non-financial competing interests that could have influenced the research findings.

Ethical Compliance

Not applicable.

Acknowledgments

The author acknowledges the financial support from the Institute of Hydrocarbon Recovery, Universiti Teknologi PETRONAS.

References

1. [△]Joseph J, Gunda NSK, Mitra SK (2013). "On-chip porous media: Porosity and permeability measurements." *Chemical Engineering Science*. 99:274–283.
2. [△]Van Geet M, Lagrou D, Swennen R (2003). "Porosity measurements of sedimentary rocks by means of microfocus X-ray computed tomography (μ CT)." *Geological Society, London, Special Publications*. 215(1):51–60.
3. [△]Nelson PH (2011). "Pore-throat sizes in sandstones, siltstones, and shales: Reply." *AAPG Bulletin*. 95(8):1448–1453. doi:10.1306/12141010159.
4. [△]Kjølstad C (2014). *Reservoir quality, diagenesis and depositional environments of Early Jurassic sandstone reservoirs located in the northern North Sea, Knarr area*. 135 p.
5. [△]Risha M (2025b). *The Effect of Clay Minerals and Microporosity on Siliciclastic Reservoirs: A Case Study from Labuan, Malaysia*. UNIVERSITI TEKNOLOGI PETRONAS. <https://utpedia.utp.edu.my/id/eprint/31239/>.
6. [△][♢]Alansari A, Salim AMA, Janjuhah HT, Bin Abd Rahman AH, Fello NM (2019). "Quantification of clay mineral microporosity and its application to water saturation and effective porosity estimation: A case study from Upper Ordovician reservoir, Libya." *Journal of Natural Gas Geoscience*. 4(3):139–150. doi:10.1016/j.jnggs.2019.04.005.
7. [△][♢]Li J, Zhong P, Yang M, Zhu F, Chen J, Liu W, Xu S (2020). "Intelligent identification of effective reservoirs based on the random forest classification model." *Journal of Hydrology*. 591:125324.
8. [△]Sahoo S, Jha MK (2017). "Pattern recognition in lithology classification: modeling using neural networks, self-organizing maps and genetic algorithms." *Hydrogeology Journal*. 25(2):311.
9. [△]Babai AMA, Ehinola OA, Gebbayin OIMFA, Ibrahim MAE (2025). "Clastic facies classification using machine learning-based algorithms: A case study from Rawat Basin, Sudan." *Energy Geoscience*. 6(1):100353.
10. [△]Hall B (2016). "Facies classification using machine learning." *The Leading Edge*. 35(10):906–909.

11. ^aRisha M, Liu P (2025). *Shoreline Prediction Models: A Review of the Evolution from Empirical to AI Machine Learning Approaches*.
12. ^a^bAli M, Changxingyue H, Wei N, Jiang R, Zhu P, Hao Z, Hussain W, Ashraf U (2024). "Optimizing seismic-based reservoir property prediction: a synthetic data-driven approach using convolutional neural networks and transfer learning with real data integration." *Artificial Intelligence Review*. 58(1):31.
13. ^aLee A-S, Enters D, Huang J-JS, Liou SYH, Zolitschka B (2022). "An automatic sediment-facies classification approach using machine learning and feature engineering." *Communications Earth & Environment*. 3(1):294.
14. ^aVashisth D, Mukerji T (2022). "Direct estimation of porosity from seismic data using rock-and wave-physics-informed neural networks." *The Leading Edge*. 41(12):840–846.
15. ^aSong Z, Li S, He S, Yuan S, Wang S (2022). "Gas-bearing prediction of tight sandstone reservoir using semi-supervised learning and transfer learning." *IEEE Geoscience and Remote Sensing Letters*. 19:1–5.
16. ^aYang Y, Zhou C, Peng J, Cai C, Tang H, Zhang J (2024). "Hybrid model-driven and data-driven method for predicting concrete creep considering uncertainty quantification." *Frontiers of Structural and Civil Engineering*. 18(10):1524–1539.
17. ^aAlolayan OS, Raymond SJ, Montgomery JB, Williams JR (2022). "Towards better shale gas production forecasting using transfer learning." *Upstream Oil and Gas Technology*. 9:100072.
18. ^aFan M, Liu S, Lu D, Gangrade S, Kao S-C (2023). "Explainable machine learning model for multi-step forecasting of reservoir inflow with uncertainty quantification." *Environmental Modelling & Software*. 170:105849.
19. ^aRoscher R, Bohn B, Duarte MF, Garcke J (2020). "Explain it to me—facing remote sensing challenges in the bio-and geosciences with explainable machine learning." *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. 3:817–824.
20. ^aPelemo-Daniels D, Stewart RR (2024). "Petrophysical property prediction from seismic inversion attributes using rock physics and machine learning: Volve Field, North Sea." *Applied Sciences*. 14(4):1345.
21. ^aEmmert-Streib F, Yang Z, Feng H, Tripathi S, Dehmer M (2020). "An introductory review of deep learning for prediction models with big data." *Frontiers in Artificial Intelligence*. 3:4.
22. ^aTariq Z, Aljawad MS, Hasan A, Murtaza M, Mohammed E, El-Husseiny A, Alarifi SA, Mahmoud M, Abdulraheem A (2021). "A systematic review of data science and machine learning applications to the oil and gas industry." *Journal of Petroleum Exploration and Production Technology*. 11(12):4339–4374. doi:10.1007/s13202-021-01302-2.

23. [△]Ajdukiewicz JM, Lander RH (2010). "Sandstone reservoir quality prediction: The state of the art." *AAPG Bulletin*. 94(8):1083–1091. doi:10.1306/intro060110.
24. [△]Xi K, Cao Y, Jahren J, Zhu R, Bjørlykke K, Haile BG, Zheng L, Hellevang H (2015). "Diagenesis and reservoir quality of the Lower Cretaceous Quantou Formation tight sandstones in the southern Songliao Basin, China." *Sedimentary Geology*. 330:90–107. doi:10.1016/j.sedgeo.2015.10.007.
25. [△]Carranza EJM, Laborte AG (2016). "Data-driven predictive modeling of mineral prospectivity using random forests: A case study in Catanduanes Island (Philippines)." *Natural Resources Research*. 25:35–50.
26. [△][‡]Rodriguez-Galiano V, Sanchez-Castillo M, Chica-Olmo M, Chica-Rivas M (2015). "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines." *Ore Geology Reviews*. 71:804–818.
27. [△]Merletti GD, Spain DR, Melick J, Armitage P, Hamman J, Shabro V, Gramin P (2016). "Integration of depositional, petrophysical, and petrographic facies for predicting permeability in tight gas reservoirs." *Interpretation*. 5(2):SE29–SE41. doi:10.1190/INT-2016-0112.1.
28. [△]Worden RH, Burley SD (2003). "Sandstone Diagenesis: The Evolution of Sand to Stone." *Sandstone Diagenesis*. 1–44. doi:10.1002/9781444304459.ch.
29. [△][‡]Taylor TR, Giles MR, Hathon LA, Diggs TN, Braunsdorf NR, Birbiglia GV, Kittridge MG, MacAulay CI, Espejo IS (2010). "Sandstone diagenesis and reservoir quality prediction: Models, myths, and reality." *AAPG Bulletin*. 94(8):1093–1132. doi:10.1306/04211009123.
30. [△]Wellmann F, Caumon G (2018). "3-D Structural geological models: Concepts, methods, and uncertainties." *Interpretation: Advances in geophysics*. Vol. 59. Elsevier. pp. 1–121.
31. [△]Yong H, Wenxiang H, Yanli Z, Bincheng G, Zhaopu G (2021). "Uncertainty, sensitivity analysis and optimization of a reservoir geological model." *Marine Georesources & Geotechnology*. 39(2):129–139.
32. [△]Zhong Z, Carr TR, Wu X, Wang G (2019). "Application of a convolutional neural network in permeability prediction: A case study in the Jacksonburg-Stringtown oil field, West Virginia, USA." *Geophysics*. 84(6):B363–B373.
33. [△]Xu C, Fu L, Lin T, Li W, Ma S (2022). "Machine learning in petrophysics: Advantages and limitations." *Artificial Intelligence in Geosciences*. 3:157–161. doi:10.1016/j.aiig.2022.11.004.
34. [△][‡]Imani M, Beikmohammadi A, Arabnia HR (2025). "Comprehensive analysis of random Forest and XGBoost performance with SMOTE, ADASYN, and GNUS under varying imbalance levels." *Technologies*. 13(3):88.
35. [△]Bentéjac C, Csörgo A, Martínez-Muñoz G (1911). "A comparative analysis of xgboost." *ArXiv Abs*. 392.

36. [△]Probst P, Wright MN, Boulesteix A (2019). "Hyperparameters and tuning strategies for random forest." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 9(3):e1301.
37. [△]Zamzam YF, Saragih TH, Herteno R, Nugrahadi DT, Huynh P-H (2024). "Comparison of CatBoost and Random Forest Methods for Lung Cancer Classification using Hyperparameter Tuning Bayesian Optimization-based." *Journal of Electronics, Electromedical Engineering, and Medical Informatics*. 6(2):125–136.
38. [△]Arigbe OD, Oyeneyin MB, Arana I, Ghazi MD (2019). "Real-time relative permeability prediction using deep learning." *Journal of Petroleum Exploration and Production Technology*. 9(2):1271–1284. doi:10.1007/s13202-018-0578-5.
39. [△]Madon M (1994). "The stratigraphy of northern Labuan, NW Sabah Basin, East Malaysia." *Bulletin of the Geological Society of Malaysia*. 36:19–30.
40. [△]Nazaruddin DA, Mansor HE, Wali SSAS (2016). "Geoheritage of labuan island." *Bulletin of the Geological Society of Malaysia*. 62:117–129. doi:10.7186/bgsm62201613.
41. [△]Risha M, Douraghi J (2021). "Impact of Clay mineral type on sandstone permeability based on field investigations: Case study on Labuan island, Malaysia." *Journal of Physics: Conference Series*. 1818(1):12091. doi:10.1088/1742-6596/1818/1/012091.
42. [△]Madon M (1997). "Sedimentological aspects of the Temburong and Belait Formations, Labuan (offshore west Sabah, Malaysia)." *Bulletin of the Geological Society of Malaysia*. 41:61–84. doi:10.7186/bgsm41199707.
43. [△]Hennig-Breitfeld J, Breitfeld HT, Hall R, BouDagher-Fadel M, Thirlwall M (2019). "A new upper Paleogene to Neogene stratigraphy for Sarawak and Labuan in northwestern Borneo: Paleogeography of the eastern Sundaland margin." *Earth-Science Reviews*. 190:1–32. doi:10.1016/j.earscirev.2018.12.006.
44. [△]Risha M, Tsegab H, Rahmani O, Douraghi J (2023). "The Impact of Clay Minerals on the Porosity Distribution of Clastic Reservoirs: A Case Study from the Labuan Island, Malaysia." *Applied Sciences (Switzerland)*. 13(6):3427. doi:10.3390/app13063427.
45. [△]Chastre C, Ludovico-Marques M (2018). "Nondestructive testing methodology to assess the conservation of historic stone buildings and monuments." In: Makhoulf ASH, Aliofkhazraei MBTH (Eds.), *Handbook of Materials Failure Analysis With Case Studies from the Construction Industries*. Butterworth-Heinemann. pp. 255–294. doi:10.1016/B978-0-08-101928-3.00013-6.
46. [△]Risha M (2025a). "AI-Assisted Thin Section Image Processing for Pore-Throat Characterization in Tight Clastic Rocks." *ArXiv Preprint ArXiv:2502.11523*.
47. [△]Wilson MD, Pittman ED (1977). "Authigenic clays in sandstones; recognition and influence on reservoir properties and paleoenvironmental analysis." *Journal of Sedimentary Research*. 47(1):3–31.

48. [△]Stracuzzi DJ, Darling MC, Peterson MG, Chen MG (2018). *Quantifying Uncertainty to Improve Decision Making in Machine Learning*. Sandia National Lab. (SNL-NM), Albuquerque, NM (United States).
49. [△]Janjuhah HT, Alansari A, Vintaned JAG (2019). "Quantification of microporosity and its effect on permeability and acoustic velocity in Miocene carbonates, Central Luconia, offshore Sarawak, Malaysia." *Journal of Petroleum Science and Engineering*. 175:108–119.
50. [△]Valdez AR, Rocha BM, Chapiro G, dos Santos RW (2020). "Uncertainty quantification and sensitivity analysis for relative permeability models of two-phase flow in porous media." *Journal of Petroleum Science and Engineering*. 192:107297.
51. [△]Li H, Wan B, Chu D, Wang R, Ma G, Fu J, Xiao Z (2023). "Progressive geological modeling and uncertainty analysis using machine learning." *ISPRS International Journal of Geo-Information*. 12(3):97.
52. [△]Tibshirani RJ, Efron B (1993). "An introduction to the bootstrap." *Monographs on Statistics and Applied Probability*. 57(1):1–436.
53. [△]Coulston JW, Blinn CE, Thomas VA, Wynne RH (2016). "Approximating prediction uncertainty for random forest regression models." *Photogrammetric Engineering & Remote Sensing*. 82(3):189–197.
54. [△]Zou C, Zhao L, Xu M, Chen Y, Geng J (2021). "Porosity prediction with uncertainty quantification from multiple seismic attributes using random forest." *Journal of Geophysical Research: Solid Earth*. 126(7):e2021JB021826.
55. [△]^bMenze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, Hamprecht FA (2009). "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data." *BMC Bioinformatics*. 10:1–16.
56. [△]Wies C, Miltenberger R, Grieser G, Jahn-Eimermacher A (2023). "Exploring the variable importance in random forests under correlations: a general concept applied to donor organ quality in post-transplant survival." *BMC Medical Research Methodology*. 23(1):209.
57. [△]Yuan X, Liu S, Feng W, Dauphin G (2023). "Feature importance ranking of random forest-based end-to-end learning algorithm." *Remote Sensing*. 15(21):5203.
58. [△]Breiman L (2001a). "Random forests." *Machine Learning*. 45(45):5–32.
59. [△]Salles T, Goncalves M, Rodrigues V, Rocha L (2015). "BROOF: Exploiting out-of-bag errors, Boosting and Random Forests for effective automated classification." *SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 353–362. doi:10.1145/2766462.2767747.

60. [△]Gorritz JM, Segovia F, Ramirez J, Ortiz A, Suckling J (2024). "Is K-fold cross validation the best model selection method for Machine Learning?" *ArXiv Preprint ArXiv:2401.16407*.
61. [△]Kärkkäinen T (2014). "On cross-validation for MLP model evaluation." *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, S+ SSPR 2014, Joensuu, Finland, August 20-22, 2014. Proceedings*. 291–300.
62. [△][♢]Bernard S, Heutte L, Adam S (2009). "Influence of hyperparameters on random forest accuracy." *Multiple Classifier Systems: 8th International Workshop, MCS 2009, Reykjavik, Iceland, June 10-12, 2009. Proceedings* 8. 171–180.
63. [△][♢]Halabaku E, Bytyçi E (2024). "Overfitting in Machine Learning: A Comparative Analysis of Decision Trees and Random Forests." *Intelligent Automation & Soft Computing*. 39(6).
64. [△]Smith HL, Biggs PJ, French NP, Smith ANH, Marshall JC (2024). "Lost in the Forest: Encoding categorical variables and the absent levels problem." *Data Mining and Knowledge Discovery*. 38(4):1889–1908. doi:10.1007/s10618-024-01019-w.
65. [△]Walter SD (2005). "The partial area under the summary ROC curve." *Statistics in Medicine*. 24(13):2025–2040. doi:10.1002/sim.2103.
66. [△]Jadhav AS (2020). "A novel weighted TPR-TNR measure to assess performance of the classifiers." *Expert Systems with Applications*. 152:113391.
67. [△]Prati RC, Batista GE, Monard MC (2011). "A survey on graphical methods for classification predictive performance evaluation." *IEEE Transactions on Knowledge and Data Engineering*. 23(11):1601–1618.
68. [△]Breiman L (2001b). "Random forests." *Machine Learning*. 45(45):5–32.
69. [△]Barreñada L, Dhiman P, Timmerman D, Boulesteix A-L, Van Calster B (2024). "Understanding overfitting in random forest for probability estimation: a visualization and simulation study." *Diagnostic and Prognostic Research*. 8(1):14.
70. [△]Chen F, Yang S, Su Z, Wang K (2016). "Effect of emissivity uncertainty on surface temperature retrieval over urban areas: Investigations based on spectral libraries." *ISPRS Journal of Photogrammetry and Remote Sensing*. 114:53–65.

Declarations

Funding: This research was funded by the YUTP grant (015LC0-060).

Potential competing interests: No potential competing interests to declare.