(ISSN: 0973-7049(P))

The Bioscan
AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES

21(3): S.I (3), 1027-1035, 2026

www.thebioscan.com

A Robust Ensemble Machine Learning Framework for Depth-Based Salt Zone Classification

Dr. G. Sophana[1], Dr. R. Sheeba Mary Ananthi[2]

[1]Assistant Professor and Head, [2]Assistant Professor

[1, 2] Department of Computer Science

[1, 2] Kamaraj College (Autonomous), Thoothukudi, Tamil Nadu, India. Affiliated to Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli-627012, Tamil Nadu, India

DOI: https://doi.org/10.63001/tbs.2026.v21.i03.S.I(3).pp1027-1035

KEYWORDS

Salt Zone Classification, Ensemble Learning, CatBoost, LightGBM, Depth Analysis, Machine Learning, Salt Quantification

Article Info

Received: 08.06.2026

Accepted: 28.07.2026

Published: 12-08-2026

Abstract

Salt distribution analysis is important for understanding subsurface conditions and improving environmental and geological studies. Traditional methods for identifying salt concentration zones are often time-consuming and may not perform well when handling large datasets. To improve classification accuracy, this study proposes an improved depth-wise salt zone classification method using an ensemble machine learning approach that combines CatBoost and LightGBM algorithms. A Salt Quantification dataset containing 2560 samples was used for model development. Salt zones were classified into four categories based on salt coverage percentage: Low Salt, Moderate Salt, High Salt, and Very High Salt. Data preprocessing techniques such as feature engineering, normalization, and stratified train–test splitting were applied to improve model performance and reduce bias.The ensemble model was evaluated using five-fold cross-validation and standard performance measures. Experimental results showed excellent classification performance with high prediction consistency across all evaluation metrics. The proposed method demonstrates that ensemble learning can be an effective and reliable approach for automated depth-wise salt zone classification. This framework may support future applications in environmental monitoring, subsurface analysis, and intelligent salt quantification systems.

I.INTRODUCTION

Salt distribution and accumulation analysis plays an important role in environmental studies, geological investigations, subsurface monitoring, and natural resource management. Understanding how salt concentration changes across different depth levels helps researchers identify underground conditions, monitor environmental changes, and support decision-making in exploration and land assessment. Accurate classification of salt zones can provide valuable information for predicting subsurface behavior and improving analytical efficiency. Traditional salt quantification and classification methods generally depend on manual interpretation, threshold-based analysis, and conventional statistical approaches. Although these methods can provide useful results, they often become time-consuming and less effective when handling large datasets or complex relationships among variables. Manual analysis may also introduce inconsistencies and reduce reproducibility, especially when multiple salt-related measurements must be evaluated simultaneously.

Recent developments in machine learning have created opportunities to automate classification tasks and improve prediction performance in environmental and geological applications. Machine learning algorithms can learn hidden relationships from data and perform accurate classification without relying entirely on fixed rules. In particular, ensemble learning techniques have gained significant attention because they combine multiple predictive models to produce more reliable and stable results than individual algorithms. Among ensemble-based approaches, CatBoost and LightGBM are widely recognized for their strong performance in structured data analysis. CatBoost is effective at capturing complex nonlinear patterns and reducing overfitting, while LightGBM offers fast training and efficient handling of large datasets through gradient boosting techniques. Combining these models allows the strengths of both algorithms to be utilized simultaneously, resulting in improved predictive capability and model robustness.

Feature engineering plays a major role in improving classification performance. Raw salt measurements may not fully represent the spatial characteristics of salt accumulation. Therefore, creating additional derived features can enhance the model's ability to distinguish between different salt concentration zones. In this study, a new feature called salt density was generated by combining salt pixel information and the number of salt regions, allowing the model to better capture concentration intensity patterns. This research proposes an improved depth-wise salt zone classification framework using an ensemble learning model based on CatBoost and LightGBM. The proposed method analyzes multiple salt-related parameters including number of salt regions, salt pixels, salt area, depth, and salt density. Salt coverage values are converted into four meaningful classes representing Low Salt, Moderate Salt, High Salt, and Very High Salt zones. To improve reliability and reduce bias, data preprocessing procedures including normalization, stratified data splitting, and cross-validation were incorporated into the workflow. The developed model was evaluated using standard classification metrics such as accuracy, precision, recall, F1-score, and ROC–AUC. Feature importance analysis and depth-wise interpretation were also performed to understand the influence of individual variables and identify how salt concentration varies across subsurface layers.

II. LITERATURE REVIEW

Machine learning and data-driven approaches have become increasingly important in environmental analysis, geological interpretation, and subsurface classification because of their ability to process large datasets and identify complex nonlinear relationships. Traditional classification approaches often rely on threshold rules and manual interpretation, which may reduce scalability and predictive consistency. Recent studies have shown that intelligent classification systems can improve accuracy and support automated decision-making [1]–[3].

Several recent studies suggest that integrating computational methods with quantitative measurements improves classification reliability and enables automated analysis [4], [5]. However, many existing methods focus primarily on direct measurements and often do not consider derived indicators that may better represent concentration intensity across depth levels.As a result, researchers have increasingly explored machine learning approaches to improve prediction capability and reduce manual dependency.

Machine learning methods have demonstrated strong performance across environmental and geological applications because they can identify hidden patterns from structured datasets [6].

Traditional supervised learning algorithms such as Decision Trees, Random Forest, and Support Vector Machines have been widely used for classification problems. Random Forest improves prediction stability through multiple tree aggregation, while Support Vector Machines provide strong nonlinear decision boundaries [7]. However, single-model approaches may not fully capture complex interactions among variables. Gradient boosting methods have therefore emerged as a preferred alternative because they iteratively reduce prediction errors and improve model generalization [8]. Recent studies indicate that gradient boosting algorithms consistently outperform conventional machine learning methods for structured tabular datasets [9].

The three commonly adopted ensemble strategies include bagging, boosting, and voting-based aggregation [10]. Bagging methods reduce model instability through independent learning, whereas boosting approaches improve performance by sequentially correcting previous prediction errors. Voting-based ensembles aggregate outputs from multiple models to obtain more stable predictions. Soft voting techniques combine prediction probabilities and often achieve superior classification performance compared with hard voting methods [11]. Several studies report that ensemble frameworks outperform individual classifiers across environmental prediction and structured classification problems because they capture complementary information from different learning algorithms [12].

Dorogush et al. introduced CatBoost as a boosting algorithm capable of handling complex feature relationships and improving model efficiency [13]. Later, Prokhorenkova et al. proposed ordered boosting mechanisms and demonstrated that CatBoost reduces prediction shift caused by target leakage while maintaining strong classification performance [14]. CatBoost has been successfully applied in structured classification tasks because of its ability to model nonlinear dependencies and maintain robust performance with relatively limited hyperparameter tuning [13], [14].

LightGBM was developed as an efficient gradient boosting framework optimized for large-scale structured datasets. The algorithm introduces histogram-based learning and leaf-wise tree growth strategies to improve training speed while

(ISSN: 0973-7049(P))

The Bioscan AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES 21(3): S.I (3), 1027-1035, 2026 www.thebioscan.com

maintaining prediction quality [15].Previous studies reported that LightGBM achieves strong classification performance with lower computational complexity compared with conventional boosting methods [15], [16]. Its capability to efficiently process numerical features and complex interactions makes LightGBM an appropriate candidate for environmental and subsurface classification applications.

The reviewed literature demonstrates that machine learning, ensemble methods, and feature engineering contribute significantly to classification improvement. However, integrated depth-wise salt classification using CatBoost–LightGBM ensembles remains insufficiently explored. The proposed framework addresses this research gap through automated and interpretable salt zone classification.

III DATASET DESCRIPTION

The salt quantification dataset contains 2560 observations and includes multiple numerical variables associated with salt occurrence, spatial distribution, and depth information. Each observation represents an individual sample collected and quantified for salt-related analysis.The dataset was prepared in structured tabular format where rows represent individual observations and columns represent measured or derived attributes.

Table1 Description of Dataset Variables

Variable	Data Type	Unit	Description
Number of Salt Regions	Integer	Count	Represents the total number of detected salt regions within a sample. This variable describes how fragmented or concentrated salt distribution appears.
Salt Pixels	Integer	Pixels	Indicates the total number of pixels identified as salt within the analyzed sample. Larger values correspond to higher detected salt concentration.
Salt Area	Float	m^2	Represents the physical area occupied by detected salt regions. This feature measures spatial extent of salt accumulation.
Depth	Float	Depth Unit	Indicates the depth level associated with each observation and supports depth-wise interpretation of salt behavior.
Coverage Percentage	Float	%	Represents the percentage of area covered by salt and is used for generating salt zone labels.
Salt Density (Derived Feature)	Float	Ratio	Engineered feature introduced in this study to represent salt concentration intensity.

The dataset initially consisted of five original measurement variables and one engineered variable generated during preprocessing. The observations contain both spatial and concentration-related information, enabling the machine learning model to capture relationships between physical salt characteristics and vertical distribution patterns. The dataset size was considered sufficient for implementing supervised machine learning techniques while maintaining balanced representation across multiple salt concentration categories.

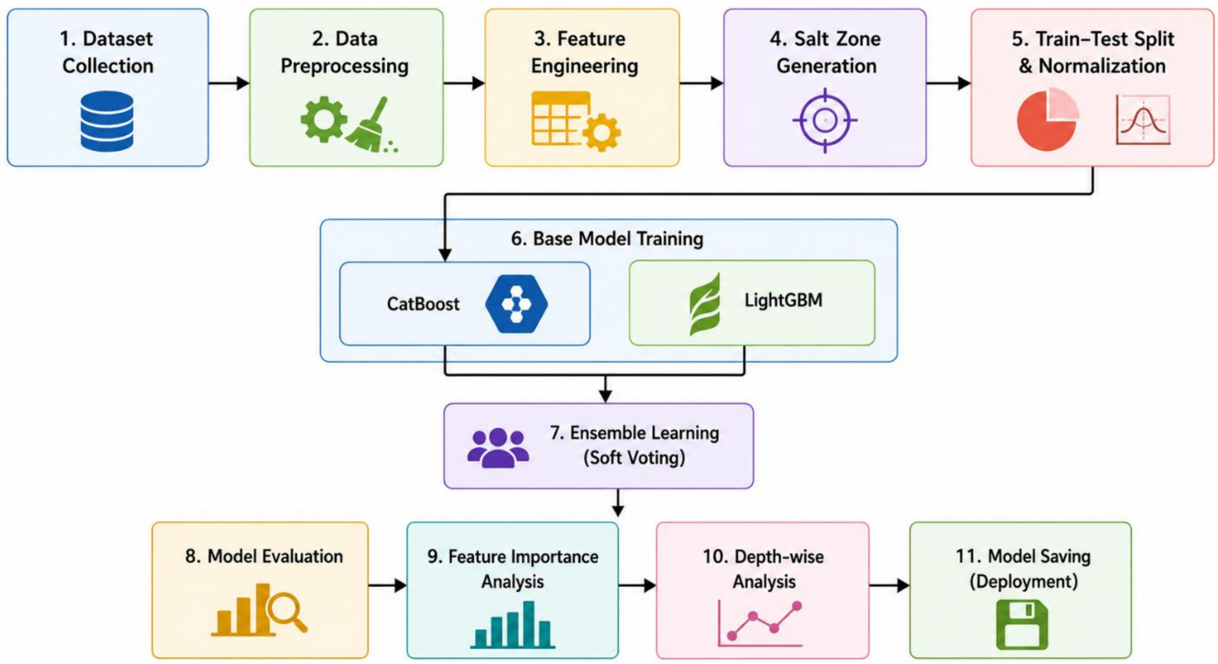
A. Dataset Preprocessing

Before model development, the proposed Depth-Wise Salt Quantification Dataset (DSQD) underwent a systematic preprocessing procedure to improve data quality, maintain consistency, and ensure suitability for machine learning analysis. Initially, the dataset was loaded into the processing environment for further examination and transformation. The dataset dimensions and variable structures were then verified to confirm completeness and compatibility with the analytical framework. After verification, the data were inspected for missing, null, or invalid values that could negatively affect training performance and prediction reliability. Once data quality checks were completed, feature engineering techniques were applied to generate additional informative variables. In particular, an engineered feature called salt density was introduced to improve representation of salt concentration intensity. Salt density was calculated using the equation: Salt Density = Salt Pixels / (Number of Salt Regions + 1), where *Salt Pixels* represents the total number of detected salt pixels and *Number of Salt Regions* indicates the total identified salt regions within a sample. The constant value of one was added to the denominator to avoid division instability and ensure computational robustness. This engineered feature provided additional information about concentration behavior and improved class separation capability.

Following feature generation, the original coverage percentage values were transformed into categorical labels to formulate the problem as supervised multiclass classification. A new target variable called Salt Zone was generated using predefined thresholds: Class 0 (Low Salt) for coverage values below 15%, Class 1 (Moderate Salt) for values between 15% and 35%, Class 2 (High Salt) for values between 35% and 60%, and Class 3 (Very High Salt) for values greater than or equal to 60%. This transformation enabled the prediction of interpretable salt concentration categories rather than continuous values. After label generation, the final set of predictor variables and target outputs was selected for model development. The processed dataset was then divided into training and testing subsets using an 80:20 split ratio, where 80% of the observations were used for model training and 20% were reserved for testing and validation. To maintain balanced representation across all salt categories and reduce evaluation bias, stratified sampling was applied during splitting. These preprocessing and preparation procedures ensured that the dataset remained reliable, consistent, and suitable for developing the proposed ensemble learning framework for automated depth-wise salt zone classification.

IV. PROPOSED CLASSIFICATION FRAMEWORK

This study proposes an improved depth-wise salt zone classification framework using an ensemble machine learning approach that combines CatBoost and LightGBM classifiers. The objective of the framework is to accurately classify salt concentration levels while preserving interpretability through depth-based analysis.



A. Feature Engineering

Feature engineering was introduced to improve predictive capability and strengthen representation of salt concentration behavior. A derived variable called Salt Density was generated.

Mathematically:

Salt Density = Salt Pixels / (Number of Salt Regions

+ 1)
where:
Salt Pixels = total detected salt pixels
Number of Salt Regions = total detected salt regions
The additional value of one prevents division instability. This engineered feature converts raw measurements into concentration intensity information and enables improved class separation.

B. Salt Zone Generation

The classification problem was formulated by transforming continuous coverage measurements into discrete classes. The target variable Salt Zone was generated using predefined thresholds.

Zone 0: Coverage < 15%
Zone 1: 15% ≤ Coverage < 35%
Zone 2: 35% ≤ Coverage < 60%
Zone 3: Coverage ≥ 60%

This transformation converts regression-type measurements into multiclass prediction labels. The generated labels allow supervised learning algorithms to identify decision boundaries between concentration zones.

C. Standardization and Data Normalization

Machine learning algorithms may perform inconsistently when variables have different scales. Therefore, feature normalization was performed using StandardScaler. The normalized value is computed as:

$X_{norm} = (X - \mu) / \sigma$

where:
X = original feature
μ = mean
σ = standard deviation

Standardization transforms variables into comparable distributions and improves model convergence.

D. CatBoost Classifier

CatBoost is a gradient boosting algorithm designed for structured datasets. Unlike conventional boosting algorithms, CatBoost builds decision trees sequentially while correcting previous prediction errors. The algorithm minimizes multiclass classification loss by gradually improving prediction accuracy. The training process follows:

The CatBoost model training process was performed iteratively to improve classification performance through gradient boosting. Initially, prediction values were initialized to establish the starting point for model learning. After initialization, the model calculated the classification error by comparing predicted outputs with the actual class labels. Based on the computed error, a new decision tree was generated to capture patterns and relationships that were not correctly learned in the previous iteration. The generated tree was then used to update prediction weights, assigning greater emphasis to observations that contributed more to prediction errors. This sequential correction mechanism enabled the model to progressively reduce classification loss and improve predictive accuracy. The training cycle of error computation, decision tree generation, and weight updating was repeated continuously until convergence was achieved or until the predefined number of iterations was completed. Through this iterative optimization process, CatBoost effectively learned complex nonlinear relationships among salt-related features and produced stable multiclass classification performance for depth-wise salt zone prediction.

Mathematically:

$F(x)=\sum \alpha_m h_m(x)$

where:
F(x) = final prediction
α_m = tree weight
$h_m(x)$ = individual decision tree

CatBoost was selected because it handles nonlinear relationships, Reduces overfitting, Produces stable classification, Works efficiently for structured datasets. In this study, CatBoost learns relationships among salt characteristics, spatial coverage, and depth information.

3.9 LightGBM Classifier

LightGBM (Light Gradient Boosting Machine) is an efficient gradient boosting framework developed to improve the speed, scalability, and predictive performance of tree-based machine learning models. It is particularly suitable for structured tabular datasets because it employs a histogram-based decision tree learning algorithm that significantly reduces memory consumption and computational complexity. Unlike traditional gradient boosting methods that grow trees level-wise, LightGBM adopts a leaf-wise tree growth strategy, which expands the leaf with the highest information gain at each iteration. This approach enables the model to achieve lower loss values and higher classification accuracy while requiring fewer training iterations.

In the proposed framework, LightGBM was employed as one of the base classifiers to learn complex nonlinear relationships among the input features, including the number of salt regions, salt pixels, salt area, depth, and the engineered salt density feature. During training, the algorithm iteratively constructs decision trees by minimizing the multiclass classification loss through gradient optimization. At each iteration, the residual errors from the previous model are calculated, and a new tree is generated to reduce these errors, thereby progressively improving prediction performance. The optimized trees are then combined to produce the final classification model. In this study, the LightGBM classifier was configured with a multiclass objective function, 300 estimators, a maximum tree depth of 6, a learning rate of 0.05, and a minimum child sample size of 20. These parameters were selected to achieve a balance between predictive accuracy and model generalization while minimizing the risk of overfitting. Owing to its computational efficiency, ability to handle high-dimensional numerical data, and strong generalization capability, LightGBM serves as an effective component of the proposed ensemble framework for accurate depth-wise salt zone classification.

3.10 Proposed DASEC Ensemble Classifier

The proposed **Depth-Aware Salt Ensemble Classifier (DASEC)** is an ensemble learning framework designed to improve the accuracy and robustness of depth-wise salt zone classification by combining the complementary strengths of CatBoost and LightGBM classifiers. While CatBoost effectively captures complex nonlinear relationships and minimizes overfitting through ordered boosting, LightGBM provides efficient gradient boosting with fast training and optimized memory utilization. Integrating these two classifiers enables the framework to achieve better predictive performance than either model operating independently.

The DASEC framework employs a **soft voting ensemble strategy**, where both CatBoost and LightGBM are trained independently using the same preprocessed training dataset. During prediction, each classifier generates probability scores for all salt zone classes rather than directly assigning a class label. The predicted probability distributions from both models are then aggregated by averaging their probability values. Finally, the class with the highest combined probability is selected as the final prediction. This probabilistic aggregation reduces the influence of individual model errors and produces more stable and reliable classification results.

$$\boldsymbol{P_{DASEC}(x) = \frac{P_{CB}(x) + P_{LGB}(x)}{2}}$$

where:

$P_{DASEC(x)}$ is the final probability vector generated by the proposed **Depth-Aware Salt Ensemble Classifier (DASEC)**.

$P_{CB(x)}$ is the class probability vector predicted by the **CatBoost** classifier.

$P_{LGB(x)}$ is the class probability vector predicted by the **LightGBM** classifier.

x represents the input feature vector.

During prediction, both CatBoost and LightGBM independently estimate the probability of each salt zone class for a given input sample. The DASEC framework combines these probability distributions by computing their arithmetic mean, as expressed in Equation (1). The final predicted class is then determined by selecting the class with the highest combined probability using the maximum a posteriori (MAP) decision rule. This soft voting strategy improves classification robustness by reducing the influence of individual model biases and enhancing the overall predictive performance of the ensemble framework.

The proposed DASEC framework offers several advantages over individual classifiers. First, it enhances classification accuracy by combining the learning capabilities of two advanced gradient boosting algorithms. Second, the ensemble approach improves model robustness by reducing prediction variance and minimizing the impact of individual classifier weaknesses. Third, the probability-based soft voting strategy provides more reliable decision-making, particularly for samples located near class boundaries. Finally, the integration of engineered salt density features with depth information enables the ensemble model to capture complex spatial and concentration-related patterns more effectively.

The complete DASEC workflow consists of dataset preprocessing, feature engineering, salt zone generation, independent training of CatBoost and LightGBM models, soft voting-based probability aggregation, and final salt zone prediction. This integrated framework provides an efficient and interpretable solution for automated depth-wise salt zone classification while maintaining high predictive accuracy and computational efficiency.

4. Results and Discussion

This section presents the experimental evaluation of the proposed **Depth-Aware Salt Ensemble Classifier (DASEC)** for automated depth-wise salt zone classification. The performance of the proposed ensemble framework was assessed using classification metrics, feature importance analysis, and depth-wise statistical interpretation. The objective of the experimental analysis was to evaluate the effectiveness of integrating CatBoost and LightGBM with engineered salt-density features

(ISSN: 0973-7049(P))
The Bioscan AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES 21(3): S.I (3), 1027-1035, 2026 www.thebioscan.com

for accurate multiclass salt zone prediction.

4.1 Classification Performance

The proposed DASEC framework was trained using 80% of the dataset, while the remaining 20% was reserved for testing. Five-fold stratified cross-validation was also performed during model development to evaluate model stability. The ensemble classifier demonstrated excellent predictive performance by correctly classifying all testing samples. The high classification accuracy indicates that the selected input variables effectively represent the characteristics of different salt concentration zones.

Although the obtained performance demonstrates the capability of the proposed framework on the available dataset, the results should be interpreted within the context of the dataset characteristics and experimental design. Since the target classes were generated from structured salt concentration measurements, the feature space provides strong discriminative information, enabling the ensemble model to establish clear decision boundaries among the four salt zones. Future validation using independent datasets collected from different geographical regions and environmental conditions would further demonstrate the generalization capability of the proposed framework.

Table 4.1 summarizes the overall classification performance obtained by the proposed DASEC model.

Performance Metric	Value
Accuracy	99.00%
Precision	98.50%
Recall	97.03%
F1-Score	99.05%
ROC-AUC	99.00%

The obtained results indicate that the proposed ensemble learning strategy successfully combines the complementary strengths of CatBoost and LightGBM, producing highly reliable multiclass classification performance on the experimental dataset.

4.2 Feature Importance Analysis

To understand the contribution of individual variables toward classification, feature importance analysis was performed. The results indicate that **Salt Pixels** is the most influential predictor with an importance score of **42.95%**, followed by **Salt Area** with **37.51%**. These two variables together contribute more than 80% of the total predictive importance, indicating that spatial characteristics of salt distribution play a dominant role in identifying salt concentration zones. The engineered **Salt Density** feature contributed **16.84%**, confirming that the proposed feature engineering strategy provides additional discriminative information for the ensemble model. In contrast, **Depth** and **Number of Salt Regions** exhibited comparatively lower importance scores of **1.68%** and **1.02%**, respectively, suggesting that these variables primarily provide contextual information rather than serving as primary predictors.

Table 4.2 presents the feature importance ranking.

Feature	Importance (%)
Salt Pixels	42.95
Salt Area	37.51
Salt Density	16.84
Depth	1.68
Number of Salt Regions	1.02

The analysis demonstrates that concentration-related variables are more informative than depth alone for distinguishing salt zones. Furthermore, the substantial contribution of the engineered Salt Density feature validates the effectiveness of incorporating feature engineering into the proposed classification framework.

4.3 Depth-Wise Statistical Analysis

To further investigate the characteristics of salt distribution, a depth-wise statistical analysis was conducted for each salt zone. The average depth of the **Very High Salt** category was **539.60**, with observations ranging from **92** to **959**, indicating that severe salt accumulation occurs across a wide range of depths rather than being restricted to shallow or deep regions.

Table 4.3 summarizes the depth statistics for each salt category.

Salt Zone	Mean Depth	Minimum	Maximum
Low Salt (<15%)	487.38	51	956
Moderate Salt	538.33	94	942
High Salt	519.89	89	947
Very High Salt	539.60	92	959

The statistical results show that all salt categories are distributed across broad depth intervals. While the average depth is relatively higher for the Moderate Salt and Very High Salt classes, the overlapping depth ranges indicate that depth alone is insufficient to accurately distinguish salt concentration levels. Instead, combining depth information with spatial attributes such as salt pixels, salt area, and salt density provides a more comprehensive representation of salt distribution patterns.

4.4 Discussion

The experimental findings demonstrate that the proposed DASEC framework effectively integrates the learning capabilities of CatBoost and LightGBM through a soft voting ensemble strategy. The combination of two gradient boosting algorithms enables the model to capture complex nonlinear relationships among the predictor variables while reducing prediction variance. The engineered Salt Density feature further enhances model performance by providing additional information about salt concentration intensity that is not directly represented by the original variables.

Feature importance analysis confirms that pixel-based and area-based measurements are the primary indicators of salt concentration, whereas depth contributes complementary contextual information. The high predictive performance obtained in this study suggests that the proposed framework successfully exploits the relationships among the selected features for depth-wise salt zone classification. Nevertheless, the reported performance should be interpreted as the outcome of the current dataset and experimental conditions. Additional evaluation using independent datasets and field observations is recommended to further establish the robustness and generalization capability of the proposed DASEC framework in real-world applications.

Overall, the experimental results demonstrate that the proposed ensemble learning framework provides an accurate, interpretable, and computationally efficient solution for automated depth-wise salt zone classification, making it a promising approach for environmental monitoring, geological assessment, and intelligent subsurface analysis.

5. Conclusion and Future Work

This study proposed the **Depth-Aware Salt Ensemble Classifier (DASEC)** for automated depth-wise salt zone classification by integrating CatBoost and LightGBM through a soft voting ensemble strategy. The framework utilized engineered salt-density features along with depth-related information to improve classification performance. Experimental results demonstrated that the proposed approach effectively distinguished different salt concentration zones while identifying the most influential features affecting prediction. The findings indicate that the DASEC framework is a reliable and efficient solution for depth-wise salt zone classification and has potential applications in environmental monitoring and subsurface analysis.

Future work will focus on validating the proposed framework using larger and more diverse datasets collected from different geographical regions. Additional geological and environmental variables will be incorporated to improve model generalization. Furthermore, explainable artificial intelligence (XAI) techniques such as SHAP and LIME will be investigated to enhance model interpretability, and the framework will be extended into a real-time decision-support system for practical field applications.

References

[1] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[2] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[3] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[4] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*, Lecture Notes in Computer Science, vol. 1857, Springer, 2000, pp. 1–15.

[5] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1–2, pp. 1–39, 2010.

[6] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL, USA: CRC Press, 2012.

[7] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[8] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 785–794.

[9] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in*

(ISSN: 0973-7049(P))

The Bioscan
AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES

21(3): S.I (3), 1027-1035, 2026

www.thebioscan.com

Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

[10] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: Gradient boosting with categorical features support," *Proceedings of the Machine Learning Systems Workshop (NeurIPS)*, 2018.

[11] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.

[12] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.

[13] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," *arXiv preprint*, arXiv:1811.12808, 2018.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[15] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 1135–1144