energies

MDPI

Article

Identification of Karst Cavities from 2D Seismic Wave Impedance Images Based on Gradient-Boosting Decision Trees Algorithms (GBDT): Case of Ordovician Fracture-Vuggy Carbonate Reservoir, Tahe Oilfield, Tarim Basin, China

Allou Koffi Franck Kouassi [1], Lin Pan [1,*], Xiao Wang [1], Zhangheng Wang [1], Alvin K. Mulashani [1,2], Faulo James [1], Mbarouk Shaame [1,3], Altaf Hussain [1], Hadi Hussain [1] and Edwin E. Nyakilla [1]

[1] Key Laboratory of Theory and Technology of Petroleum Exploration and Development in Hubei Province, China University of Geosciences, Wuhan 430074, China
[2] Department of Geoscience and Mining Technology, College of Engineering and Technology, Mbeya University of Science and Technology, Mbeya P.O. Box 131, Tanzania
[3] Department of Petroleum and Energy Engineering, College of Earth Sciences and Engineering, The University of Dodoma, Dodoma P.O. Box 259, Tanzania
* Correspondence: panlin@cug.edu.cn

Abstract: The precise characterization of geological bodies in fracture-vuggy carbonates is challenging due to their high complexity and heterogeneous distribution. This study aims to present the hybrid of Visual Geometry Group 16 (VGG-16) pre-trained by Gradient-Boosting Decision Tree (GBDT) models as a novel approach for predicting and generating karst cavities with high accuracy on various scales based on uncertainty assessment from a small dataset. Seismic wave impedance images were used as input data. Their manual interpretation was used to build GBDT classifiers for Light Gradient-Boosting Machine (LightGBM) and Unbiased Boosting with Categorical Features (CatBoost) for predicting the karst cavities and unconformities. The results show that the LightGBM was the best GBDT classifier, which performed excellently in karst cavity interpretation, giving an F1-score between 0.87 and 0.94 and a micro-G-Mean ranging from 0.92 to 0.96. Furthermore, the LightGBM performed better in cave prediction than Linear Regression (LR) and Multilayer Perceptron (MLP). The prediction of karst cavities according to the LightGBM model was performed well according to the uncertainty quantification. Therefore, the hybrid VGG16 and GBDT algorithms can be implemented as an improved approach for efficiently identifying geological features within similar reservoirs worldwide.

Keywords: fracture-vuggy carbonate reservoir; karst cavities; gradient-boosting decision trees (GBDT); Visual Geometry Group 16 pre-trained (VGG-16); uncertainty; Tahe oilfield

Citation: Kouassi, A.K.F.; Pan, L.; Wang, X.; Wang, Z.; Mulashani, A.K.; James, F.; Shaame, M.; Hussain, A.; Hussain, H.; Nyakilla, E.E. Identification of Karst Cavities from 2D Seismic Wave Impedance Images Based on Gradient-Boosting Decision Trees Algorithms (GBDT): Case of Ordovician Fracture-Vuggy Carbonate Reservoir, Tahe Oilfield, Tarim Basin, China. *Energies* **2023**, *16*, 643. https://doi.org/10.3390/en16020643

Academic Editors: Weichao Yan and Huaimin Dong

Received: 1 November 2022
Revised: 14 December 2022
Accepted: 28 December 2022
Published: 5 January 2023

1. Introduction

Carbonate reservoirs have emerged as one of the crucial targets for oil and gas production in many basins worldwide [1,2]. The discoveries of various enormous oil fields, such as the Tahe oilfield and associated reservoirs in the Tarim Basin, with hydrocarbon resources exceeding 10^6 tons, have turned into a significant exploration prospect in China [3–6]. The large and main production layer reservoirs in the Tahe oilfield are the Yingshan formations and Yijianfang formation of the middle-lower Ordovician fractured-vuggy carbonate [7,8]. According to the structure and sediment lithology, these reservoirs have firm heterogeneity with an ultra-low to low matrix permeability and porosity [9–11]. The fractured-vuggy system consists of many complex geological features (karst, caves, vugs,

pores, and fractures) varying in spatial distribution, geometrical morphology, forms, scales, and connectivity [3,12,13].

In line with this, production from these carbonate reservoirs depends strongly on caves or the combination of both fractures and caves for their storage ability [14–16]. A realistic geological model has to carefully consider these geologic bodies for an excellent reservoir-scale numerical conceptual model. The karst cavities in the Tahe oilfield are this study's primary object, including paleocaves, vugs, and pores. This is because intrinsic heterogeneities complicate observations and make estimation difficult. As a result, identifying karst cavities as reservoirs has usually been challenging in the petroleum field. Therefore, an advanced method that reduces complexity in karst cavity interpretation is required for better estimation and analysis.

Many researchers have shown the indispensability of geophysical approaches in detecting and analyzing various caves and evaluating reserves from seismic attributes [11,16–18]. These include the usage of microgravity, magnetic, electrical resistivity tomography (ERT), induced polarization (IP) methods, and ground-penetrating radar (GPR) in several studies to investigate heterogeneities of the karst features, such as karstified zone limits, filled or unfilled paleocaves, voids, and sinkholes [19–21]. In addition, some authors recommend the combination of two or more geophysical methods to delineate the target better, reduce uncertainty, and avoid misinterpretations of the target [22,23].

Cave facies extraction is regarded as a segmentation problem that is easily handled by supervised machine learning methods. Supervised learning methods can provide the best results by learning from predefined labels as output data and unlabeled patterns as input data through seismic data. Seismic diffraction data, seismic reflection, conventional seismic images, and acoustic impedance images are some of the best tools that efficiently facilitate understanding the structure of caves by mapping their 2D or 3D spatial distribution in fractured-vuggy reservoirs [23–26]. However, few studies have been conducted on paleocave characterization using supervised learning methods.

Attempts at unsupervised machine learning-based paleocave characterization include the usage of a proposed sparsity constraint inverse spectral decomposition [27], the usage of adopted waveform clusters, spectral decomposition, geological constraints, and fuzzy C-Means clusters to characterize and classify the fracture-cavity paleo-channel reservoirs [28,29] and the application of a democratic neural network association to predict lithofacies filling caves of paleokarst [30]. Similarly, studies have proposed neural network models for paleocave identification. These include the usage of a supervised convolutional neural network (CNN) model to generate unlabeled training images automatically and labeled images for collapsed paleokarst feature characterization using 3-D seismic images to avoid wasting time and to solve the problem of lacking training datasets [31], using Bayesian encoder-decoder network from a synthetic seismic dataset to characterize the paleocaves [32] and the application of mask region-convolutional neural network method to extract carbonate cavities from a digital outcrop profile automatically [33]. Only manually interpreting the karst cavities from conventional seismic images or simulation datasets predefined from seismic attributes ensures that the label data is reliable [32,33]. However, it takes much time to prepare many label images by hand for training a 2D or 3D CNN model. One of the best ways to eliminate the need for an extensive training dataset is to use a non-neural network algorithm.

Gradient-boosting decision trees (GBDT) algorithms are non-neural network machine-learning methods developed for regression and classification problems. It is a new and powerful machine-learning technique with a solid capacity to learn and deal with different scales of features with nonlinear decision boundaries [34]. Among other GBDT methods, Light Gradient Boosting Machine (LightGBM), Unbiased Boosting with Categorical Features (CatBoost), and Extreme Gradient Boosting Machine (XGBoost) are used to predict almost every domain of petroleum, including lithologic and facies prediction [35–40], petrophysical parameters prediction [40–43], production forecast [40,44], well drilling [45], prediction of pseudo density log [46], and enhanced oil recovery [47].

Nevertheless, GBDT algorithms, which do not require many datasets, have not yet been applied to identify karst cavities.

As with all machine learning, the GBDT algorithm cannot automatically learn features from training image data [48]. In contrast to deep learning models, they require features that are currently accessible. VGG-16, on the other hand, is well-known not only for its classification performance as a robust deep CNN-based model but also for its powerful extractor-enhanced classification feature [48]. However, VGG-16 needs more than 100 images or many data in order to make a good classification. The hybridization of VGG-16 and GBDT resulted in a parallel processing neural network that extracted high-level features that improved the target accuracy. The proposed method is thus worth establishing.

The primary purpose of this study is to present a novel hybrid method of Visual Geometry Group 16 (VGG-16) pre-trained and GBDT to characterize the spatial geometry of karst cavities based on uncertainty reduction from the 2D seismic wave impedance small dataset of Tahe oilfield, Tarim Basin. First, the hybrid VGG-16 with the GBDT method for unconformities and karst cavities prediction was proposed. Secondly, the best approaches for karst cavities prediction were identified. Lastly, the benefits of coupling VGG-16 and GBDT models for estimating karst cavities were assessed regarding uncertainty. The proposed model in this work can be used in any domain to see intricate features like the geological features of the paleo-karst reservoir and produce more precise subsurface models that could be important for petroleum, hydrogeology, geophysics, geological hazards, and engineering research. In other words, the proposed model can be used in any field to address the image segmentation issue.

2. Geological Setting

The Tahe oilfield is one of China's most crucial carbonate oil and gas fields. It is in the northern Tarim Basin and covers more than 3200 km^2 [26,49]. The study area is situated in the Tahe oilfield's northwestern part, near the Luntai fault that runs east–west (Figure 1a). It is an important area (approximately 600 km^2) with huge exploration potential [10]. The Lower-Middle Ordovician carbonates, karstified by the Early Hercynian, cover its significant reserve. The Middle Ordovician (Yingshan-Yijianfang) Formations are the area of interest for our research. Geologically modified by the impact of several geodynamic phenomena such as tectonic activities (Caledonian, Hercynian, Indo-Yanshanian, and Himalayan movements) and associated with three episodes of karstification (I, II, and III) and series of erosions, Ordovician carbonate reservoirs are covered by ultradeep faulted karst and multiple complexes of karst cavity (vugs, paleocave, and pores) in the Tahe oilfield, developing a significant potential resource favorable to the hydrocarbon migration and accumulation of hydrocarbons [23,26,50,51]. Early Hercynian karstification was the most important in making the complex paleokarst systems, including karst cavities. These asymmetrical and tubular cavities were partly or entirely by different sediments formed in the lower-middle Ordovician Formation's vadose zones [23,51,52].

Following the distance between the top of the Ordovician and the Shuangfeng limestone, three karst geomorphic units—the karst highlands, I-level karst slope, and II-level karst slope—formed in the Tahe region. They were produced under the influence of the Sangtamu waterproof layer, paleo-topographic alterations, and karst hydrodynamic conditions in the overlay region during the genesis of Early Hercynian karstification [10,51].

Middle Ordovician strata generally formed a shallow platform environment and were mainly composed of grainstone and dolomitic limestone [7,17,50] (Figure 2).

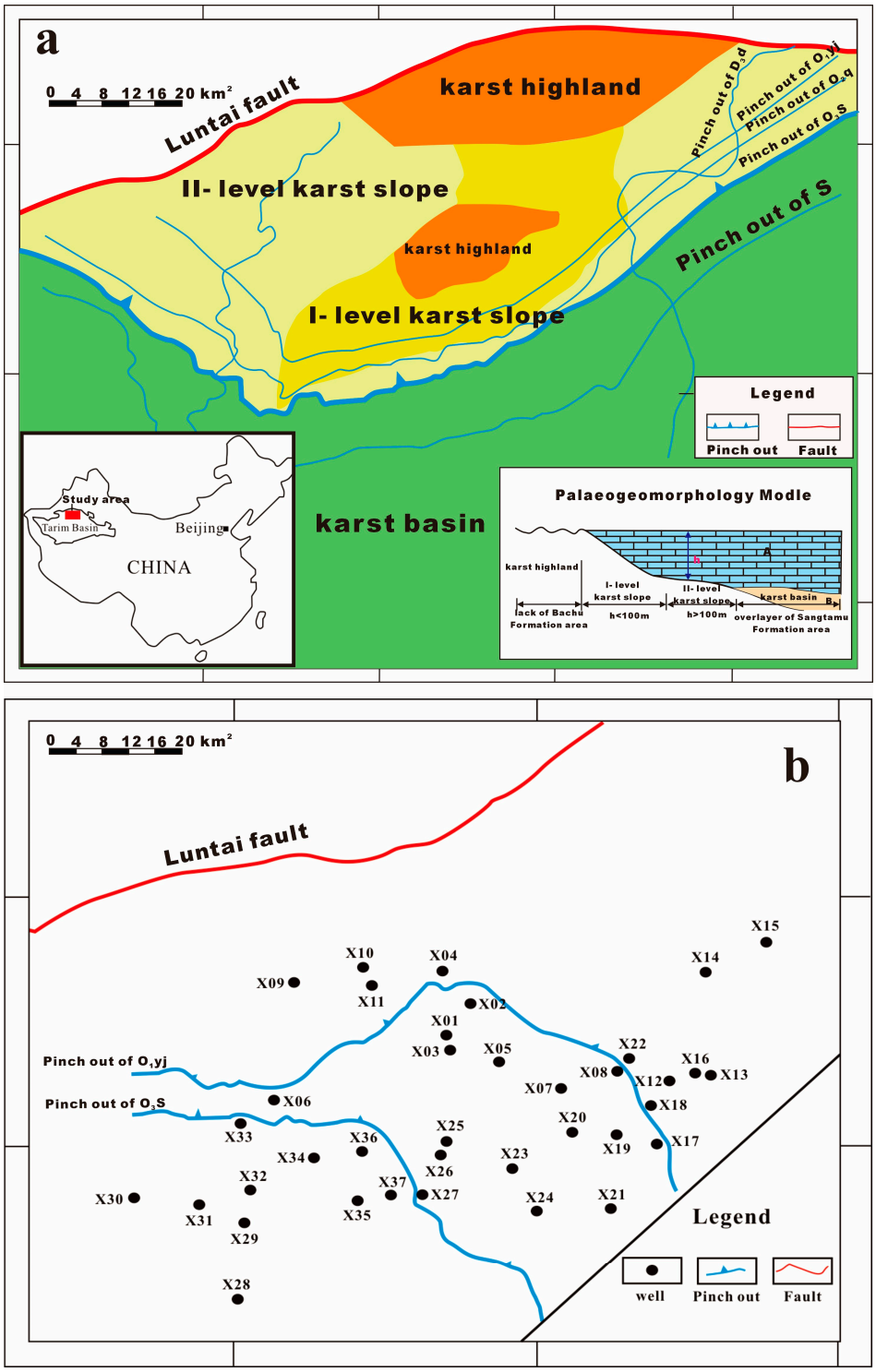

Figure 1. Early Hercynian karst paleo-topography of Tahe oilfield (**a**) and study well in the study area (**b**) (modified from [10]).

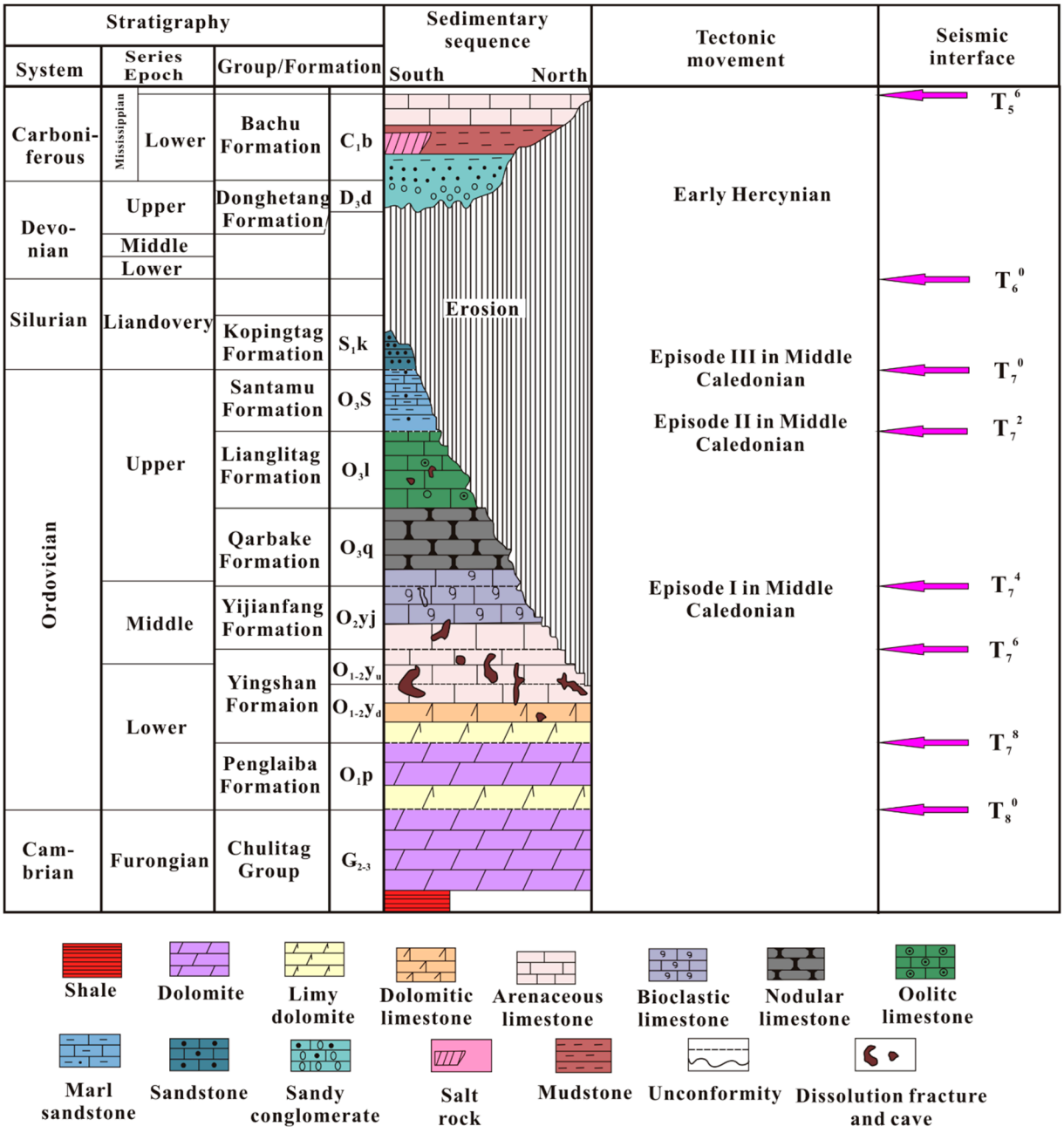

Figure 2. Stratigraphic column of the Tahe oilfield, from the Upper Cambrian to the Lower Carboniferous (modified from [53,54]).

3. Material and Methods

3.1. Seismic Wave Impedance Image Datasets

The dataset used in this study was composed of seismic wave impedance images of the Lower to Middle Ordovician carbonate reservoirs. The dataset consisted of seventy and fifty-five vertical images in color (RGB channels) with a dimension of 1247 × 552 pixels oriented north–south (N–S) and 976 × 552 pixels oriented east–west (E–W) (Figure 3), respectively. Eight unlabeled images of fifty-five images and nine of seventy were randomly selected as input data and interpreted manually using AUTOCAD software as output data. Then the assigned images were all annotated using ENVI software.

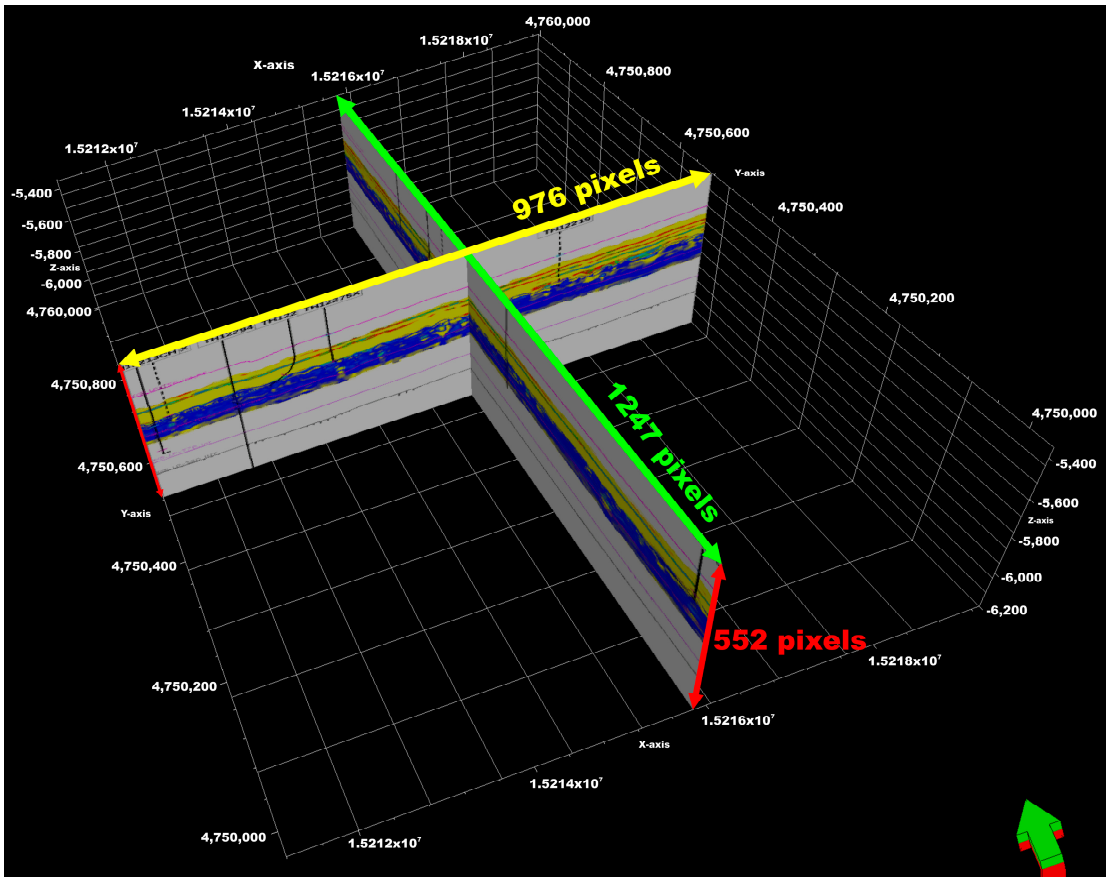

Figure 3. Image introducing dataset sizes and orientation. The green arrow indicates the direction north.

3.2. Gradient-Boosting Decision Tree (GBDT)

3.2.1. Principle of GBDT Algorithm

Gradient-boosting methods (GBM) are machine learning techniques that gradually integrate several weak learners, such as simple decision trees, to develop a complex and robust model with greater accuracy depending on each sub-model's residual error [45,55,56]. The data weights are adjusted for each weak learner individually, and the weightings of the decision trees determine their accuracy in each iteration tree [57]. The residual yields from an iteration become the input for the following decision tree [34,58]. In other words, the fundamental idea of the GBM is to create a new sub-model to offset the residual error produced by the preceding sub-model [45]. The gradient-boosting decision tree (GBDT) model is a classifier produced from an ensemble of decision trees that integrates a series of weak base learners used in gradient-boosting splits to address the overfitting problem [59,60].

In this current study, the input training set, consisting of extracted features based on VGG-16 pre-trained and labelled target variable set (uniformities and karst cavities), is $\{(X_i, y_i)\}_{i=1}^{n}$, and n is the number of samples of dataset. As shown in Figure 4, during the procedure of the GBDT model, each weak model predicts an output composed of a residual error and the desired output in every iteration. The first residual error (h_1) can be determined by this following equation:

$$y_1 = h_1 + \gamma f_1(X, \theta_1) \quad (1)$$

where θ the parameter of specific classifiers that controls the structure of tree, γ designates the weight of each weak learn, X is the input variables, and f_1 denotes the first weak learn, $f_1(X, \theta_1)$ defines the output of the first regression tree.

At the $m-1^{th}$ time of iteration, the weak learner is $f_{m-1}(X,\theta_{m-1})$, and the loss function becomes $L(y,f_{m-1}(X,\theta_{m-1}))$. The residual error generally decreases when the number of regression trees increase. At the final round, where the smallest residual is reached, the weak learner is h_m, and the loss function values can be estimated from the following equation as:

$$L(h_m,f_m(X,\theta_m)=L(h,f_{m-1}(X,\theta_{m-1})+h_m(X,\theta_{m-1})))\quad(2)$$

where *m is* the number of regression trees. The loss function is the primary factor in determining whether the model is suitable for the solved issue. The loss function must be reduced to improve accuracy and keep the model stable, necessitating using a negative gradient to maintain residual fitting. The GBDT model employs the regression algorithm to identify the optimal θ_j and construct $f(X,\theta_j)$ at the j^{th} step to minimize the objective function (Equation (3)) [58]. The objective function is calculated following this equation:

$$L=\sum l(\hat{y}_i,y_i)=\sum_i l[\hat{y}_i^{i-1}+\gamma f_j(X_i,\theta_j),y_i]\quad(3)$$

where l is the loss function. At the m^{th} iteration, the negative gradient value of the loss function is given by [59]:

$$r_{ti}=-[\tfrac{\partial L(y_i,f(X_i,\theta_i))}{\partial f(X_i)}]_{f(X)=f_{t-1}(X),}\quad(4)$$

Or

$$r_{ti}=-[\tfrac{\partial L\left(y_i,f(X_i,\theta_i)\right)}{\partial f(X_i)}]_{f(X)=f_{t-1}(X)}=y-f(X_i)\quad(5)$$

When the error loss function is the squared error between the predicted value and the truth output value. The final output of the model is the sum of all the weak models' predictions. This would be written as [58]:

$$\hat{y}^m=\hat{y}^{m-1}+\gamma f_m(X,\theta_m)=\gamma\sum_{j=1}^m f_j\,(X,\theta_j)\quad(6)$$

where θ_j is the parameter controlling the structure of j^{th} tree, and $\hat{y}$ is the prediction of the j^{th} regression tree; $f(X,\theta_j)$ is the output of the j^{th} regression tree.

LightGBM and CatBoost are derived from the GBDT algorithm used in this study.

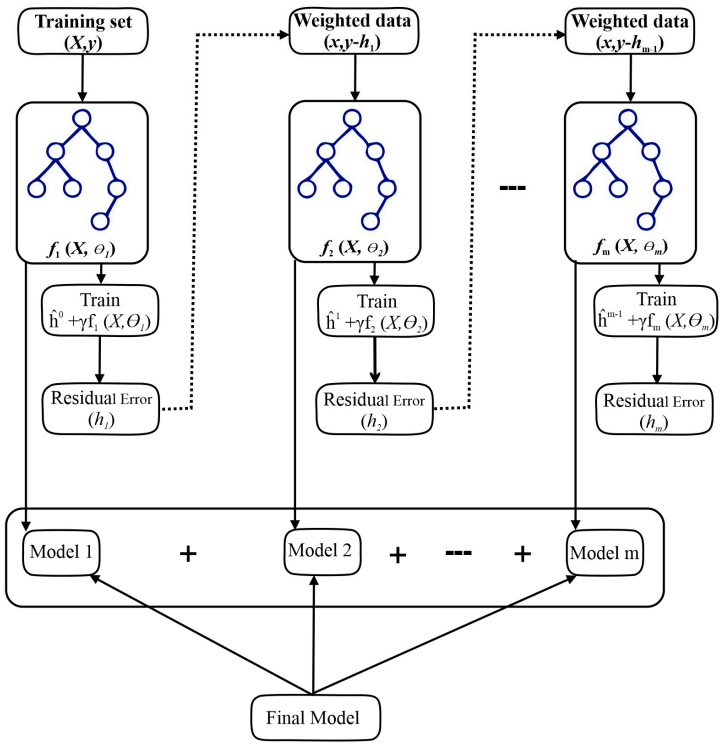

Figure 4. Schematic diagram of the GBRT model.

3.2.2. LightGBM

The first GBDT-based approaches, such as XGBoost, could not efficiently process and evaluate the information obtained when the datasets contained large amounts of data in terms of computational time [61]. Microsoft proposed the LightGBM algorithm based on the boosting regression algorithm. LightGBM uses three principal strategies to ensure that a practice is completed quickly, efficiently, and precisely. Firstly, LightGBM used leaf-wise tree growth to build its decision tree [62]. However, to guarantee training efficiency and prevent overfitting, the depth of the tree and the minimum data of each leaf node were both regulated by LightGBM.

Histogram-based techniques can aid in lowering loss, speeding up training, and reducing memory utilization [63]. Secondly, LightGBM splits the internal nodes using the gradient-based one-side sampling (GOSS) method based on variance gain. GOSS decreases the number of instances with modest gradients before computing information gain, allowing it to sample enhanced data [64]. The histogram-based approach takes longer to compute than GOSS. Finally, LightGBM employs exclusive feature bundling (EFB) techniques to reduce the size of input features and speed up the training process without sacrificing accuracy. More information regarding GOSS and EFB theories can be found [64,65].

3.2.3. CatBoost

CatBoost is a GBDT method proposed by [66,67]. It has certain peculiarities compared to other GBDT models. CatBoost can use datasets with categorical features for training and testing, unlike other GBMs. During any machine preprocessing step, the categorical features are usually converted into numerical features [68,69]. CatBoost can convert features to numbers thanks to greedy target-based statistics (Greedy TBS) [67]. Secondly, CatBoost uses a novel method termed "ordered boosting", which efficiently adapts gradient-boosting methods to tackle the target leak problem [66]. At last, CatBoost can handle a small dataset well. We frequently use stochastic permutations for the training data in CatBoost, which improves the algorithm's robustness. If we have a dataset $D = (X_i, Y)$ and a permutation $\sigma = (\sigma_1 \ldots, \sigma_n)$, then the substituted $x_{\sigma p,k}$ is [67,70]:

$$\frac{\sum_{j=1}^{p-1}\left[x_{\sigma j,k} = x_{\sigma p,k}\right]Y_{\sigma j} + aP}{\sum_{j=1}^{p-1}\left[x_{\sigma j,k} = x_{\sigma p,k}\right] + a} \quad (7)$$

where p is a prior value, and a is the weight of the initial value. This strategy aids in the reduction of noise generated by the low-frequency category.

3.3. Study Workflow

The key steps for further explaining the different phases of methodology are feature extraction, data pre-processing for hybrid VGG-16 and GBDT model construction, performance evaluation metrics, and uncertainty assessment. Figure 5 shows a brief description of the workflow of the undertaken study.

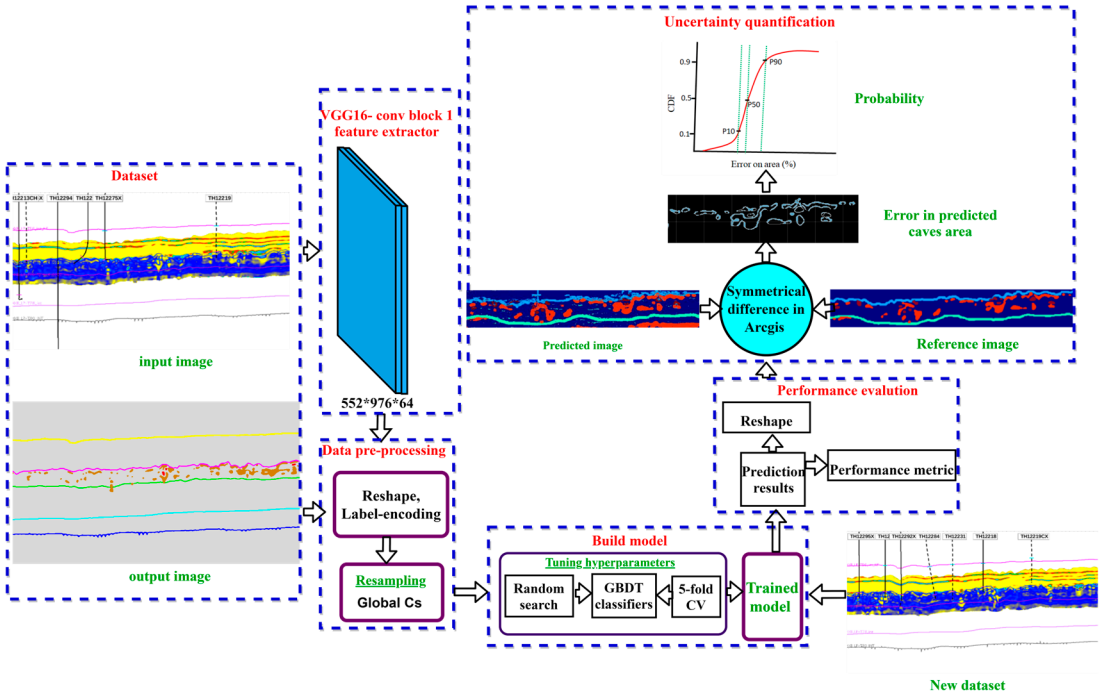

Figure 5. Study workflow.

3.3.1. Feature Extraction

The Visual Geometry Group of the University of Oxford built an efficient group of deep convolutional neural networks (CNNs), namely VGGNet, consisting of ResNet101, VGG-19, DenseNet201 ImageNet, and VGG-16, for feature extraction and classification problems [71–74]. In this paper, we modified the VGG-16 architecture according to the input image sizes and applied its first block (VGG-16-Conv block 1) as feature extractors. The architecture of our pre-trained VGG-16 model consists of 5 max-pooling layers and 13 convolutional layers (Figure 6). The main goal of feature extraction is to extract pixel values from images to be used by any machine learning model.

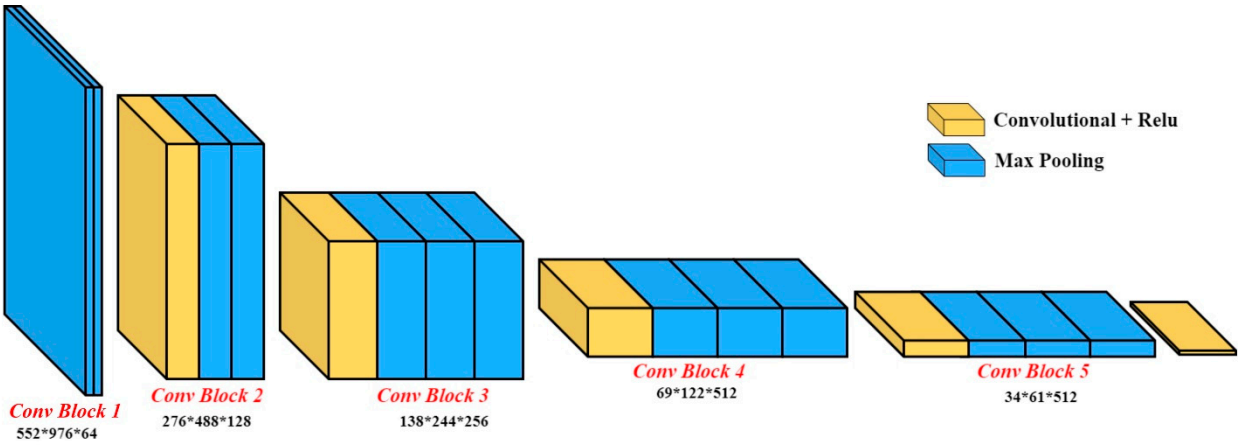

Figure 6. Pre-trained VGG-16 architecture.

3.3.2. Data Pre-Processing for Hybrid VGG-16-GBDT Model Construction

After the pre-trained step, extracted features from seismic wave impedance and labeled images were reshaped to make them further match one another. The labeled data were then classified (one-hot encoded). This study used the global CS method to treat the class imbalance problems to avoid a biased result. Global CS is a resampling algorithm

sourced from the Python module package to overcome multi-class imbalanced datasets [75]. It duplicates all samples equally for each class to achieve the majority class size. The new balanced dataset was finally used as training data by different machine learning classifiers: hybrid VGG-16 and GBDT models. Before building GBDT models, hyperparameter models were individually tuned by the randomized search model and evaluated using the 5-fold cross-validation (CV) method.

3.3.3. Performance Evaluation Metrics

Statistical measurements, such as the F-1 score, the weight-geometric mean (weight-G-Mean), the micro-geometric mean (micro-G-Mean), and the multi-class area under the receiver operating characteristic curve (multi-class AUC-ROC) were used to figure out the effectiveness of each classifier.

i. F-1 score

Recall refers to a classifier's ability to detect available samples, whereas precision refers to the accuracy of recognizing relevant samples [76] on class ω_i; precision and recall are expressed as follows:

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \tag{8}$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \tag{9}$$

TP_i is the number of true positives, FP_i is the number of false positives, and FN_i is the number of false negatives. $F-1$ score is a precision and recall harmonic mean [77]. High, low, or null values of the $F-1$ score depend on the increase, decrease, or null values of precision and recall. For multi-class datasets, computing $F-1$ score values of all the classes is defined as [76]:

$$F-1\ score = \frac{1}{c}\sum_{i=1}^{c}\frac{2.Recall_i.Precision_i}{Recall_i+Precision_i} \tag{10}$$

ii. G-Mean

Geometric-Mean (G-Mean), like the $F-1$ score, is a proper single metric for unbalanced data issues and is defined as:

$$G-mean = (\prod_{i=1}^{w} Recall_i)^{\frac{1}{c}} \tag{11}$$

Two statistical measurements of the $G-Mean$, called the *micro-G-Mean* and the *weight-G-Mean*, can help assess a multi-class classifier. The *Micro-G-Mean* approach adds up the system's false positives, true positives, and false negatives for various sets and uses them to calculate statistics. The $weight-G-Mean$ determines each label's statistics and multiplies them by their weight before being averaged.

iii. ROC-AUC

AUC-ROC curves are diagnostic plots used for classification problems that summarize a model's ability to discriminate several classes. [78] defines the AUC as follows:

$$AUC = \int_0^1 TPR(t_i)dFPR(t_i) \tag{12}$$

$TPR(t_i)$ is the true positive rate, and $FRP(t_i)$ represents the false positive rate; [79] divided ROC-AUC values into five classes, as described in Table 1. The macro-average AUC and micro-average ROC curves are necessary in the case of multiple classifiers to evaluate the prediction performance [80].

Table 1. Discrimination accuracy of ROC–AUC (adapted from [79–81]).

AUC Values	Interpretation
0.5–0.6	Not discrimination
0.6–0.7	Poor discrimination
0.7–0.8	Fair discrimination
0.8–0.9	Good discrimination
0.9–1	Excellent discrimination

3.3.4. Uncertainty Assessment

In this study, uncertainty is linked to errors in predicted area caves by the machine learning models. The errors in the area, which is the difference area between the predicted caves and references, were computed by ArcGIS software through an overlay analysis tool called the symmetrical difference. Based on the probability density (PDF) and cumulative probability density function (CDF), the frequency of these incorrectly classified cave areas was used to assess the uncertainty prediction. CDF was regarded as the primary statistical uncertainty probability method.

4. Results

4.1. Training Models Results

The models were built during training in two cases, A and B. The models in each case were constructed following the study workflow described in Figure 4. The models in case A were constructed from seismic wave impedance images of 1247 × 552 size, and those in case B were built at a 976 × 552 size (Figure 3). In each case, three images were used for training. For testing models, six images-oriented N–S and eight images-oriented E–W were used in case A, and five images-oriented N–S and five images-oriented E–W were used in case B.

The range of selected hyperparameters, their importance, the best optimal values, and the results of the five-fold CV evaluation of each GBDT classifier based on F1-scores are mentioned in supplementary Tables S1 and S2, respectively. The training results in cases A and B were, respectively, 0.99 and 0.98 for the LightGBM model and 0.96 and 0.86 for the CatBoost model. Table S3 reveals that LightGBM wastes less time than the CatBoost model in case A, but the opposite is true in case B.

4.2. Model Performance

This part compares the performance of GBDT models in terms of cave and unconformity prediction. Table 2 shows the results of each model. In both cases, A and B, the F1-score, micro-G-Mean, and weight-G-Mean of LightGBM have the highest values, indicating that it performed better than CatBoost. In case A, the mean values of the F1-score, micro-G-Mean, and weight-G-mean are, respectively, 0.83, 0.90, and 0.89 for CatBoost and 0.89, 0.94, and 0.91 for LightGBM. In case B, the average of F1-score and micro-G-Mean were, respectively, 0.85 and 0.91 for CatBoost and 0.92 and 0.96 for LightGBM. However, based on the first (Q1) and third (Q3) quartile values of the weight-G-Mean, CatBoost is introduced as having the best performance with 0.84 and 0.89, respectively, in case B. Indeed, both models statistically achieved significant performance, although the results show that LightGBM classified better than CatBoost except in Q1 and Q3.

Table 2. Statistic parameters of used GBDT models.

Models	Parameters	Case A			Case B		
		F1-Score	Weight-G-Mean	Micro-G-Mean	F1-Score	Weight-G-Mean	Micro-G-Mean
CatBoost	Min	0.82	0.80	0.89	0.83	0.73	0.90
	Q1	0.82	0.89	0.89	0.84	0.84	0.91

	Median	0.83	0.91	0.90	0.85	0.88	0.91
	Mean	0.83	0.89	0.90	0.85	0.86	0.91
	Q3	0.85	0.92	0.91	0.86	0.89	0.92
	Max	0.86	0.92	0.92	0.87	0.91	0.93
	Min	0.87	0.80	0.92	0.90	0.71	0.94
	Q1	0.88	0.94	0.93	0.92	0.78	0.95
LightGBM	Median	0.89	0.93	0.94	0.93	0.85	0.96
	Mean	0.89	0.91	0.94	0.92	0.83	0.96
	Q3	0.90	0.94	0.94	0.93	0.88	0.96
	Max	0.91	0.94	0.94	0.94	0.91	0.96

Figures 7 and 8 show the classification performance of each classifier to discriminate between each feature in both cases A and B. In case A, all classifiers failed to extract classes 6 and 7 from images 7480, 7490, and 7500 (Figure 7) and from images 2520 and 2530 in case B (Figure 8). Classes 6 and 7 are big and small caves, respectively. In these mentioned samples, the maximum values of the ROC-AUC curves of classes 6 and 7 were 0.86 and 0.75, respectively, in case A and 0.89 and 0.73 in case B. The minimum values of the ROC curves are observed with LightGBM, estimated to be 0.62 for class 7 in case A and estimated to be 0.65 for class 6 and 0.35 for class 7 in case B. Moreover, in other samples of each case, all classifiers performed well with all classes. Minimum micro- and macro-average ROC curves were, respectively, 0.96 and 0.88 in case A and 0.97 and 0.76 in case B. From a statistical point of view based on the ROC-AUC curves, CatBoost may perform better than LightGBM. This is not corroborated with previous performance metrics that carried us to further visualization analysis.

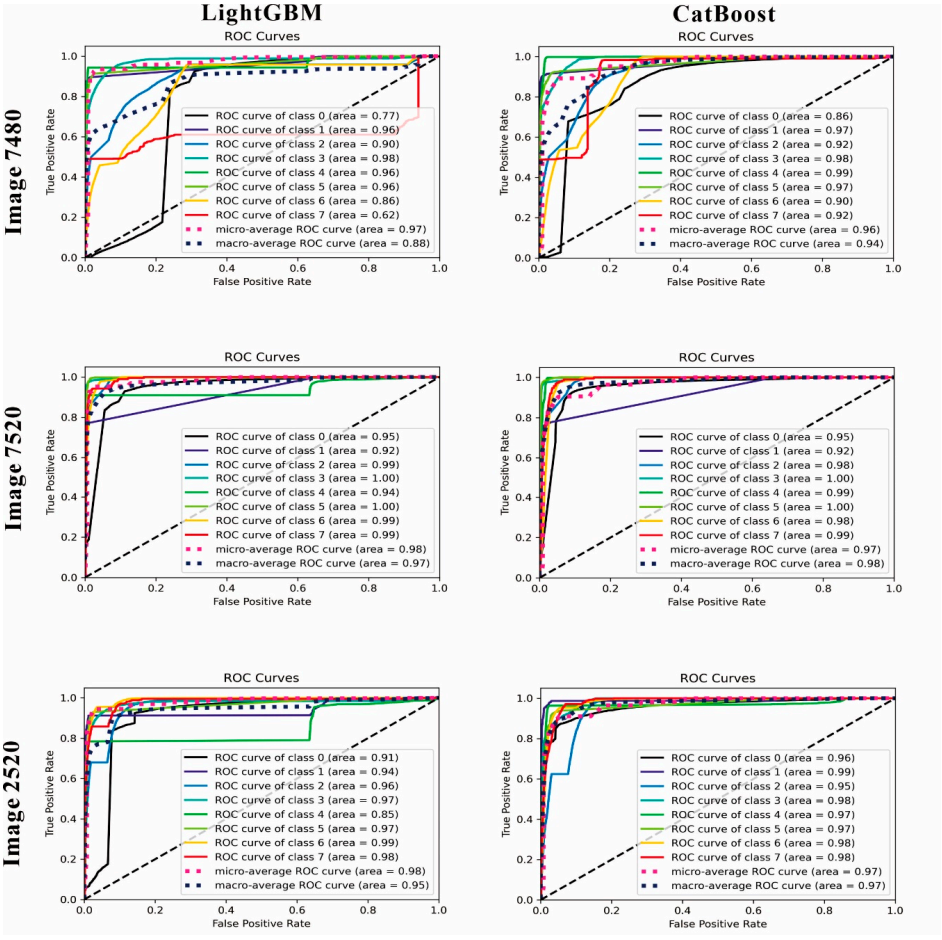

Figure 7. Multi-class ROC-AUC curves of GBDT classifiers in case A.

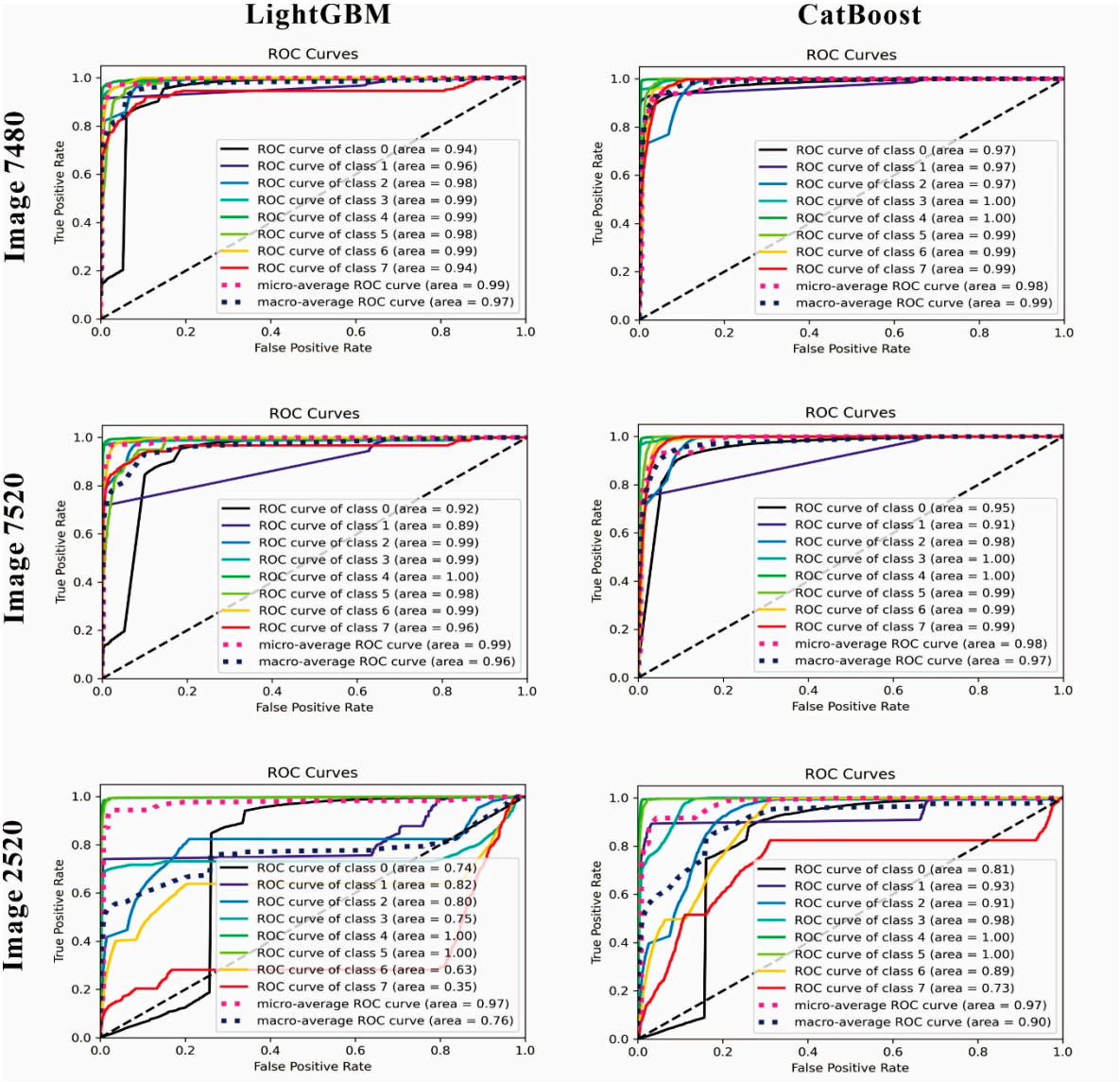

Figure 8. Multi-class ROC-AUC curves of GBDT classifiers in Case B.

4.3. Comparison of Models' Capacity for Generating Features

Figures 9 and 10 show the spatial distribution of caves and unconformities generated by GBDT classifiers in cases A and B. They quickly permitted the visual comparison of the classification performance of each used model. To evaluate the performance of each model, we focused on the predicted karst features that are between T74 and T76. In case A, Figure 9a,b,e,f demonstrated that LightGBM reproduced unconformities and caves while drastically reducing noise. However, CatBoost predicted caves and boundaries accurately but with some noise (Figure 9c,d). GBDT models could properly generate the caves and unconformities in case B with very little noise (Figure 10), except T74. Moreover, a profound observation of each predicted image showed that LightGBM performed better than CatBoost. It was deemed the best GBDT model due to its ability to learn and predict the similar main features of reference images of any size and orientation. Therefore, after comparing the predicted images for cases A and B (Figure 11), the images made by the LightGBM model for case B were used to figure out how much uncertainty there was.

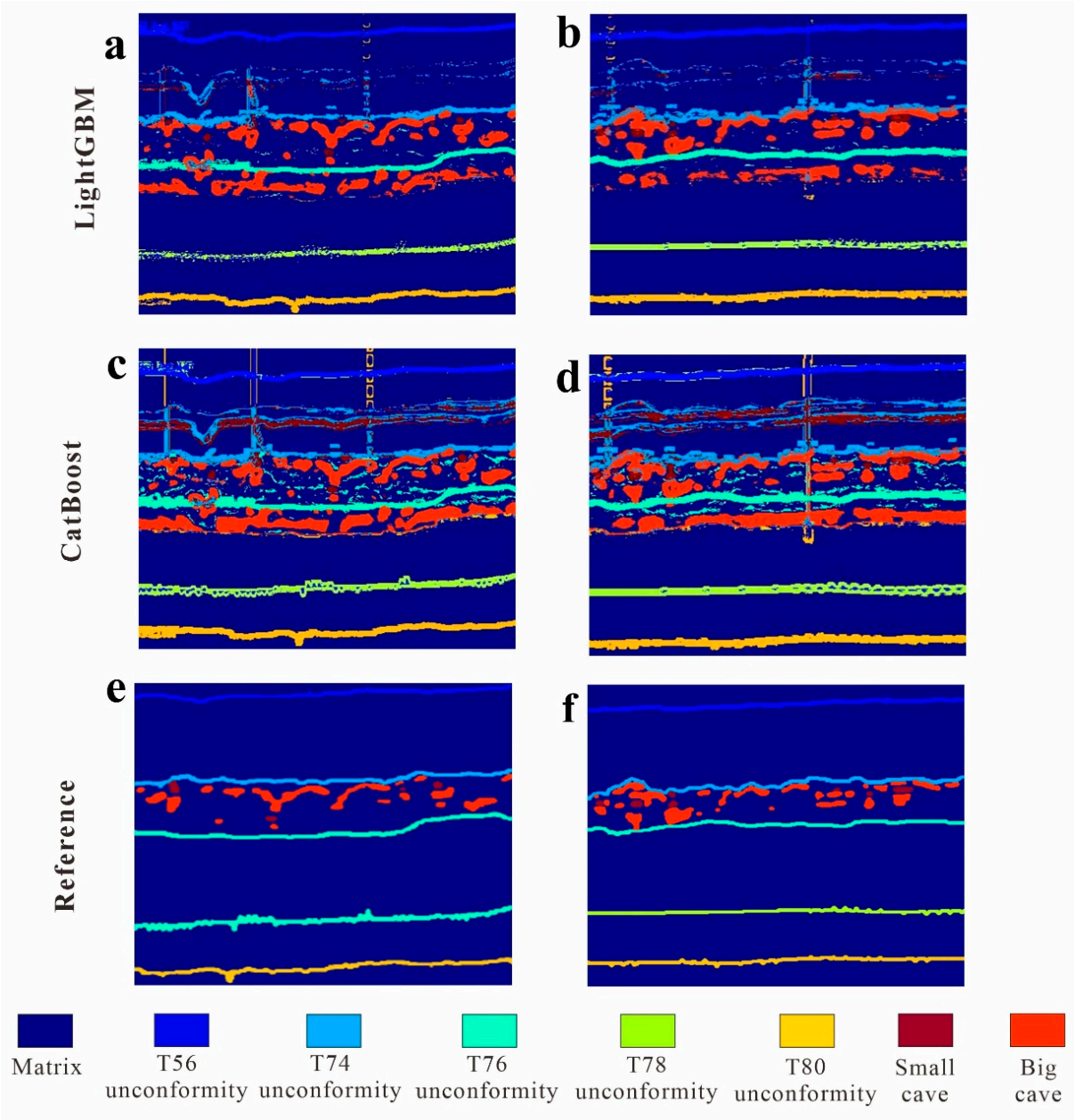

Figure 9. Comparison of images predicted by LightGBM (**a**,**b**), CatBoost (**c**,**d**), and reference images 7480 (**e**) and 2530 (**f**) in case A.

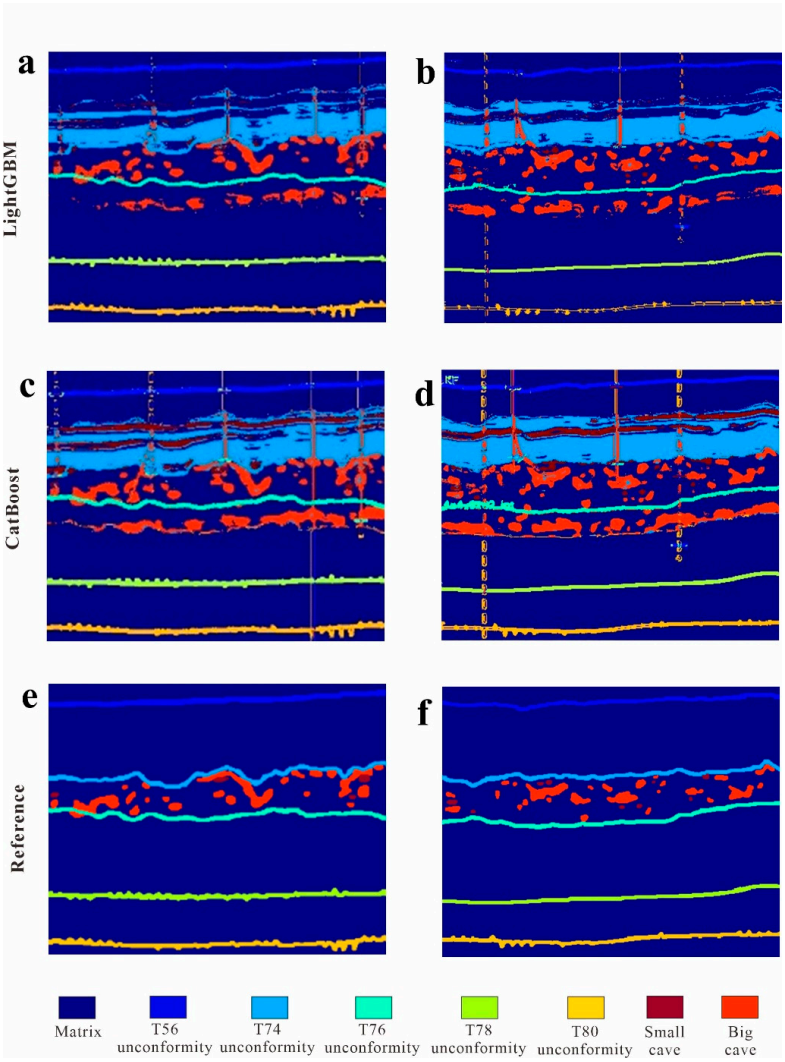

Figure 10. Comparison of images predicted by LightGBM (**a**,**b**), CatBoost (**c**,**d**), and reference images 2520 (**e**), 7490 (**f**) in case B.

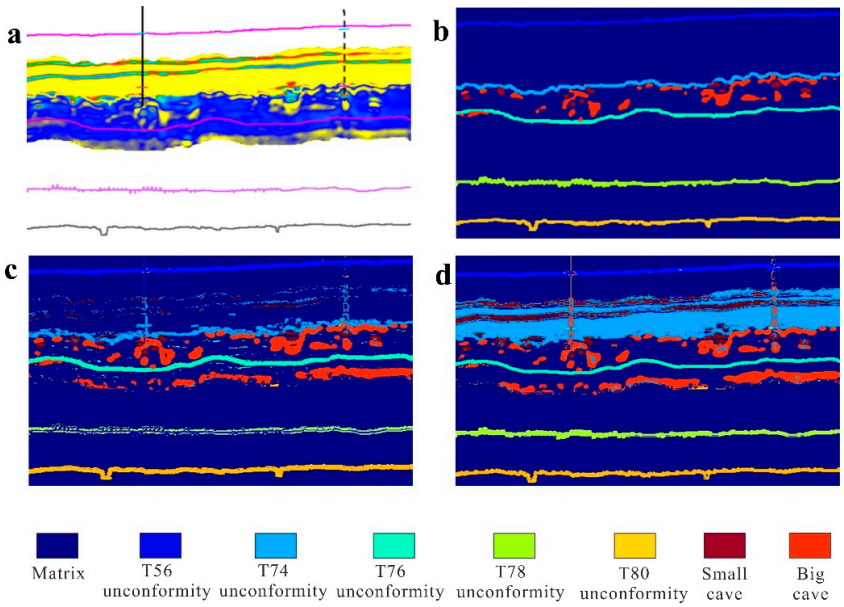

Figure 11. Comparison of the reference image 2550 (**b**) with the predicted images by LightGBM in cases A (**c**) and B (**d**) from image 2550 (**a**).

4.4. Uncertainty Assessment for the Spatial Distribution of Geometry Caves

4.4.1. *Orientation and Channel Connectivity of Caves Geometry*

The green and deep pink polygons shown in Figures 12a and S1a are the shapes of small and big predicted caves, respectively. Polygons filled in red and dark orange are small and big reference caves. These figures resulted in a good performance of the LightGBM model because the channel geometries and the spatial association of caves are correctly and reasonably segmented. Even though, at some locations, the channel connection of big caves was overestimated (missing to preserve their discontinuity), the results provided a good match between the facies channel of the estimated caves and the reference ones in terms of orientation and connectivity.

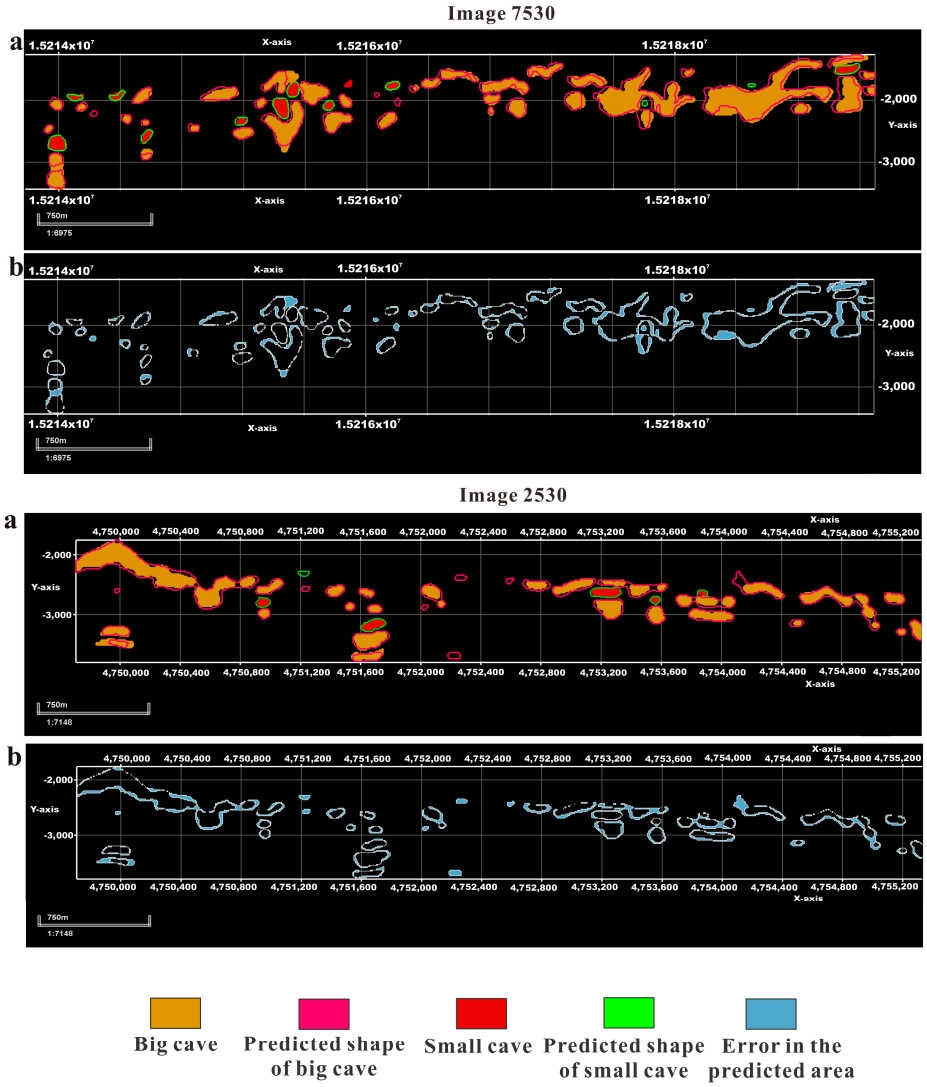

Figure 12. Predicted cave facies by the LightGBM model compared to references (**a**) and visualization of under- or overestimated cave area (**b**).

4.4.2. Uncertainty Quantification

Figures 12a and S1a have confirmed that LightGBM handled the location and spatial distribution of caves well, whatever their size. It also provided a perfect segmentation of cave patterns, similar to the references. However, predicted caves gradually overestimate, underestimate, or are slightly higher than references. As a result, quantifying uncertainty in cave prediction becomes an essential task in this study.

Figures 12b and S1b highlight errors in the predicted cave area by the LightGBM model. PDF and CDF are shown in Figure 13. The P10, P50, and P90 of the CDF curves are 63, 34, and 19%, respectively. This indicates that the uncertainty in cave prediction provided by LightGBM is highly favorable. In other words, the probability of the model overestimating or underestimating the area of cave geometry ranges from 19% to 34%.

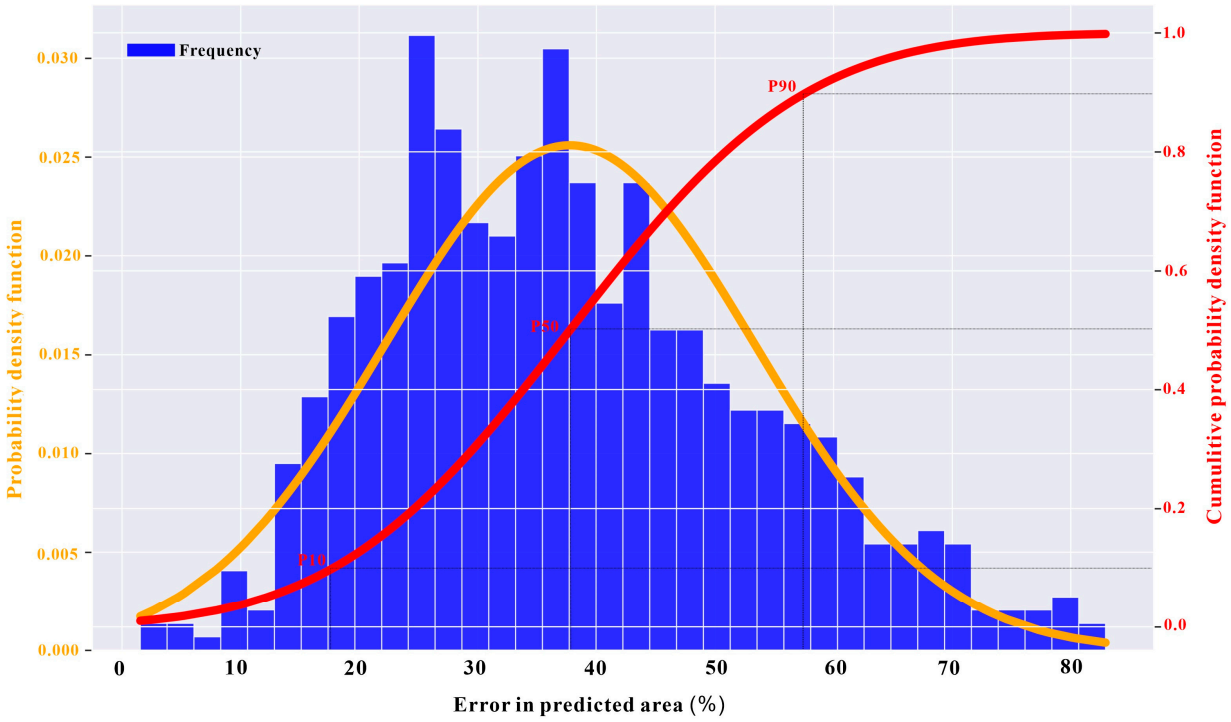

Figure 13. Histogram diagram, probability density, and cumulative distribution curves of error in the predicted cave area by LightGBM.

5. Discussion

This section aims to evaluate the performance of the LightGBM and CatBoost models compared to the logistic regression (LR) and multilayer perceptron (MLP) models. We can assess how trustworthy our suggested approaches are using the input data.

LR is a machine learning method generally used for classification or regression based on a probability function. In the case of binomial or multi-class classification, LR trains dependent variables with weights to predict categorical variables as output using the Bernoulli probability function (Equation (7)) [82]. The values of these output variables can only range from 0 to 1.

$$P_i = \frac{e^{\alpha+\beta_i x_i}}{1+e^{\alpha+\beta_i x_i}} \quad (13)$$

where P_i is the probability for a specific value of xi, α is the intercept, β designs the regression coefficient, and e is the base of the natural logarithm base.

The MLP is a type of popular feedforward neural network with a simple structure composed of three interconnected layers. The input layer, its first layer, receives the input neurons. The second layer is made up of one or more hidden layers of neurons that learn

and compute the data that was received. Finally, the output layer predicts or classifies the output [83–85].

The specific procedure to build and train LR and MLP models was the same as that for the GBDT models described in Figure 4. The range of their selected hyperparameters, their importance, the best optimal values, and the results of the 5-fold CV evaluation of both classifiers are mentioned in Tables S1 and S2, respectively. Table S3 illustrates how long it takes to train and build these models.

5.1. Model Performances Comparison

Both boxplot diagrams in Figure 14 exhibit the classification performance of all models employed in this study to predict caves and unconformities. In case A, the minimum values of the F1-score, micro-G-Mean, and weight-G-Mean of both the LightGBM and CatBoost models are more significant than the maximum values of the LR and MLP models, indicating that the GBDT models performed better than both the MLP and MLP models. The F1-score, micro-G-Mean, and weight-G-Mean are estimated at 0.83, 0.9, and 0.88 for CatBoost and 0.94, 0.96, and 0.88 for LightGBM. However, the maximum values of the F1-score, micro-G-Mean, and weight-G-Mean were 0.76, 0.87, and 0.86 for MLP and 0.78, 0.87, and 0.82 for LR. Figure 15a–c shows that, compared to MLP and LR, LightGBM with the F1-score, macro-G-Mean, and micro-G-Mean, all greater than 0.9, correctly classified unconformities and karst cavity features in more than 100% of samples and did better than CatBoost with the F1-score, macro-G-Mean, and micro-G-Mean, all greater than 0.9. In case B, Figures 14b and 15d,e showed that only the LightGBM model performed better than MLP. In brief, although LightGBM performance may decrease in particular samples, as indicated by the weight-G-Mean values in Figure 15f, it is the best robustness model. It wastes less computation time (Table S3).

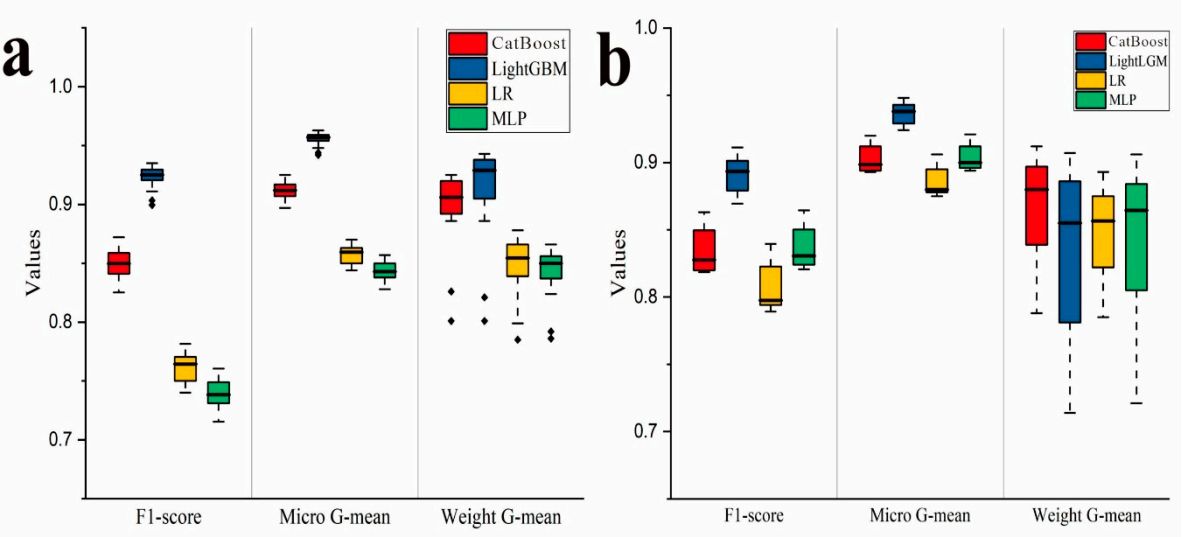

Figure 14. Boxplot diagram of the prediction results of the used models in cases A (**a**) and B (**b**).

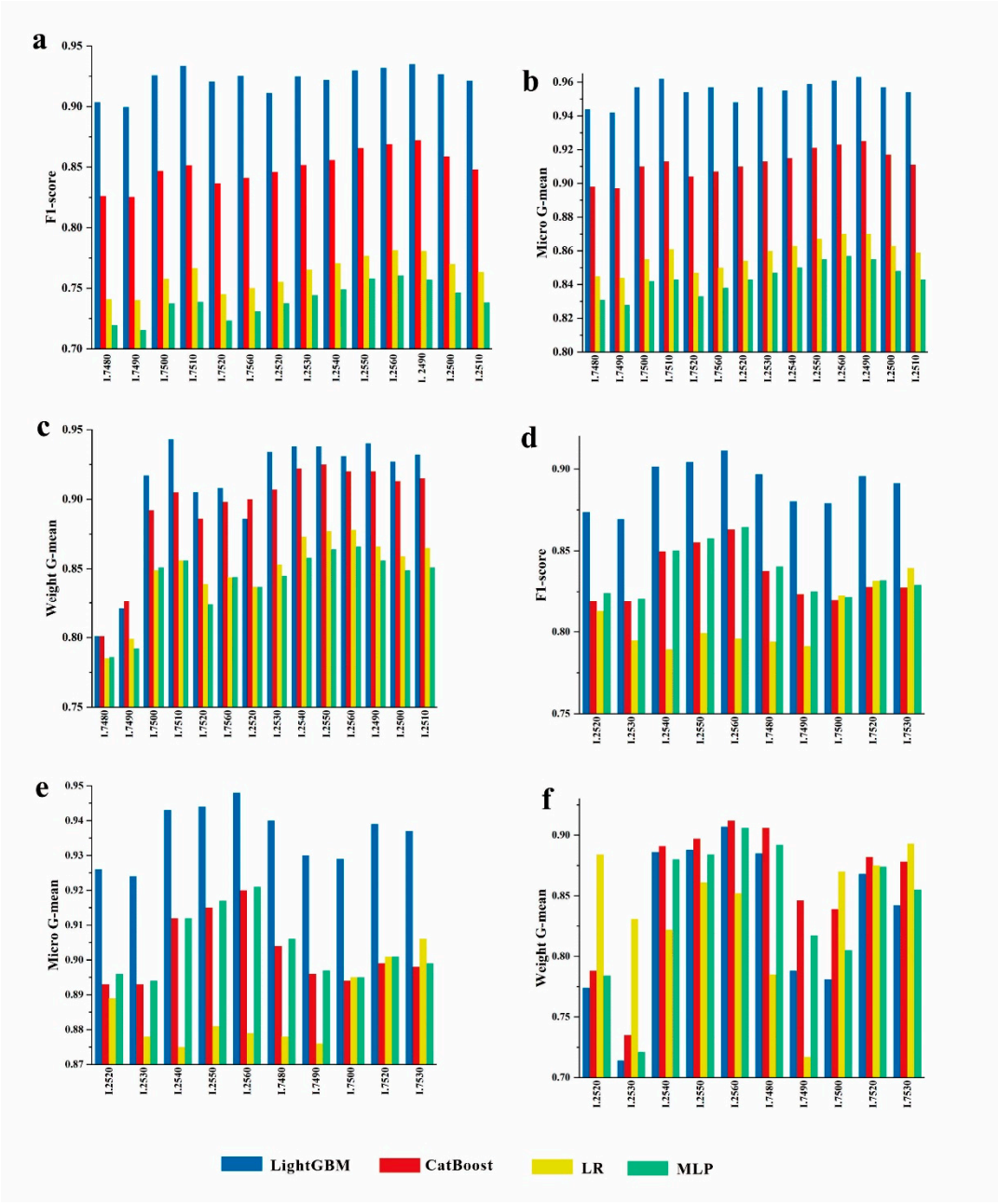

Figure 15. Comparison of model performances through the F1-score, macro-G-Mean, and micro-G-Mean in Case A (**a**–**c**) and Case B (**d**–**f**) using bar charts.

Figures 16 and 17 show that MLP and LR have similar performance according to ROC-AUC curves. Both MLP and LR could not ideally discriminate between classes 6 and 7 of images 7480, 7490, and 7500 in case A and 2520 and 2530 images in case B. Therefore, MLP and LR also performed admirably as GDBT models, caves, and uniformities in other images.

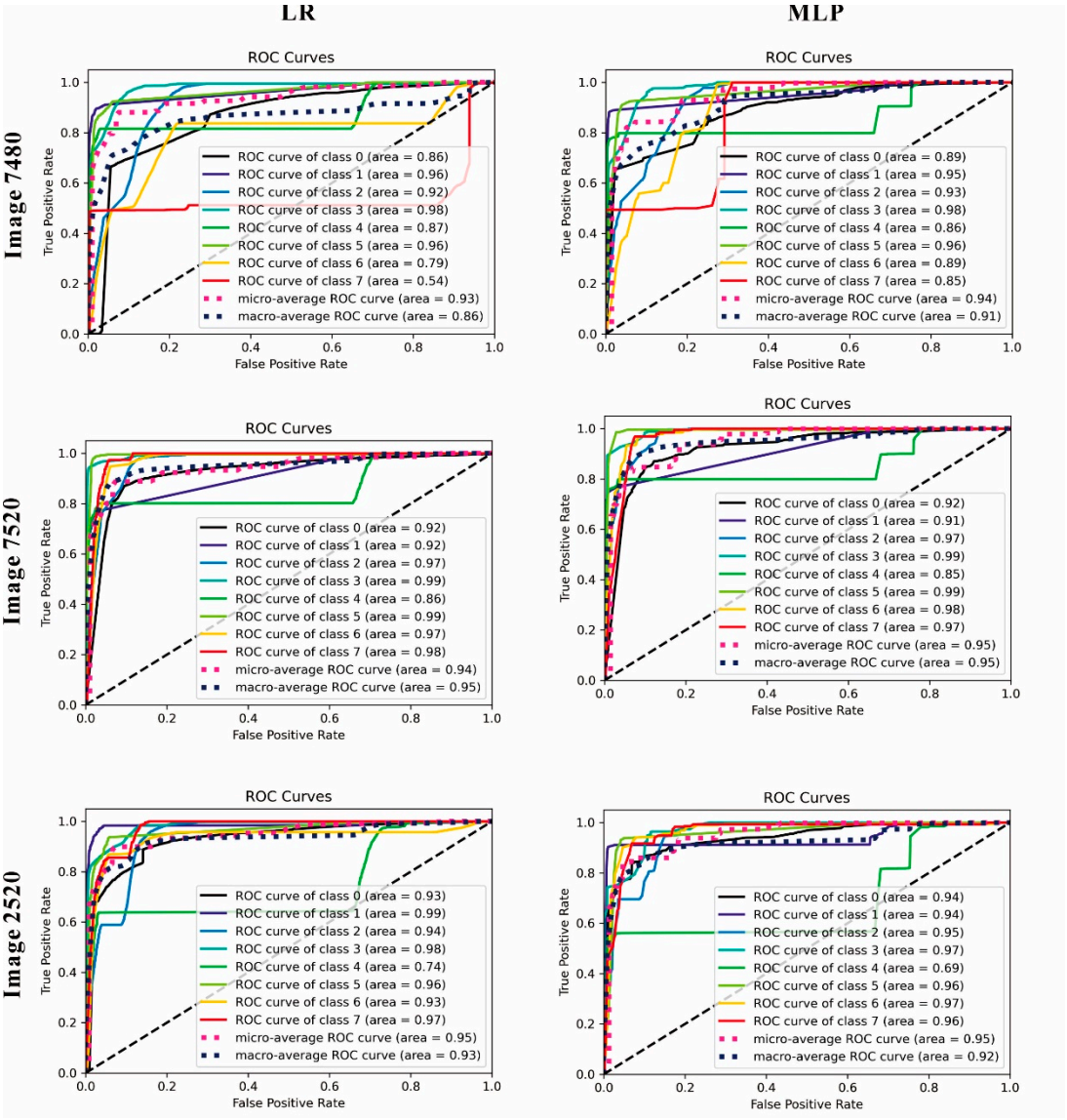

Figure 16. ROC-AUC analysis of LR and MLP classifiers in case A.

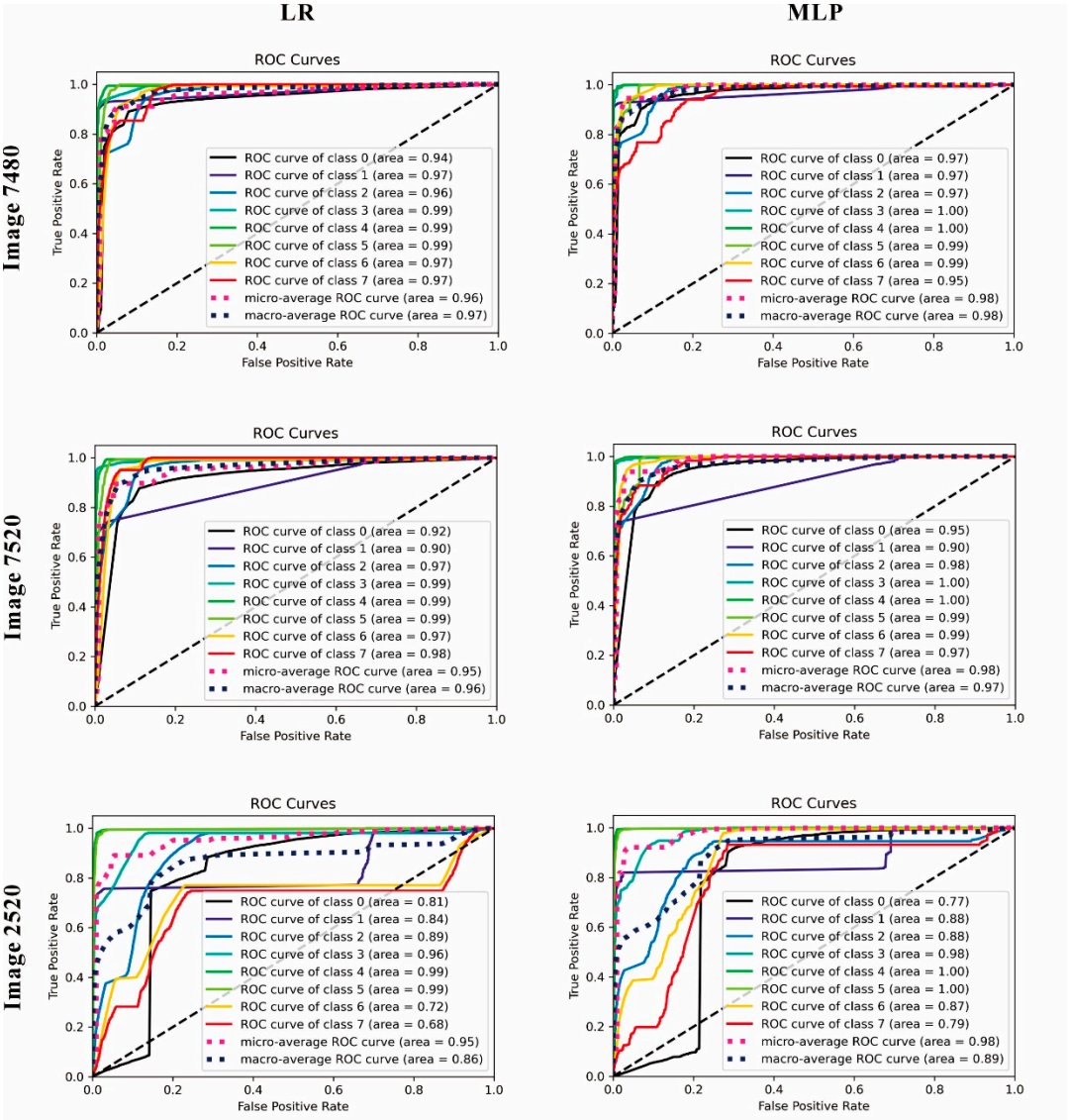

Figure 17. ROC-AUC analysis of LR and MLP classifiers in case B.

5.2. Comparison of Models' Capacity for Generating Features

In case A, Figure 18a,c,e showed that MLP and LR failed to differentiate T76 and matrix and strongly overestimated small caves and T74, causing a lot of noise. MLP as GBDT is able to preserve unconformities and different caves with very few noises (Figures 9, 10 and 18b,f), in contrast to LR (Figure 18d), which slightly failed to predict caves well in Case B. Therefore, LightGBM remained the best classifier in terms of cave prediction.

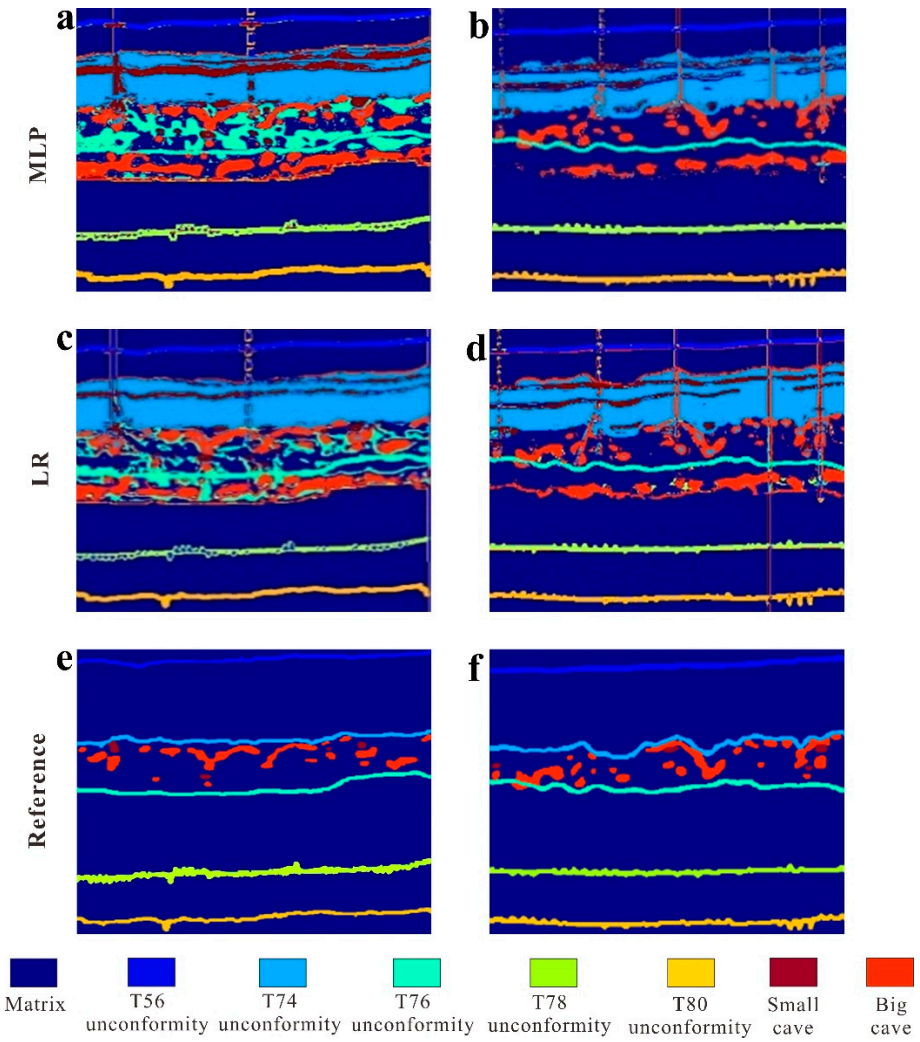

Figure 18. Comparison of images predicted by MLP (**a**,**b**), LR (**c**,**d**), and reference images (**e**,**f**).

5.3. Uncertainty and Suggestion

According to the CDF analysis, the uncertainty prediction in Figure 12 seems relatively great or acceptable. This uncertainty may be due to model misclassification; noises during the process of shape extraction in ENVI software, especially during the image smoothing step; and calculating coordinates through linear functions in Excel. These sources of errors were mentioned by [44]. In addition, our proposed method consumes less computational time. Our approach accurately provided an excellent result based on previous study results because it better classified the small and big karst cavity geometries and unconformities from a few datasets. Other studies using seismic approaches to extract and characterize paleocaves did not quantify the uncertainty that can reduce the risk of making a decision [11,13,16,86]. The workflow of the present work applied a straightforward method for evaluating predicted karst cavity uncertainty based on statistical evaluation, compared to the Bayesian deep learning used [32] based on observation. It is not easy to compare the performance of our proposed methods to the Bayesian encoder–decoder network. For leading risk assessment in exploration and development, the proposed methods can help with the 3D geological model.

As mentioned before, the methods proposed by [31–33] for identifying paleocaves based on CNN require large datasets for training. Our proposed workflow can provide an encouraging result using very few datasets. For the fast and accurate prediction of caves in future work, the researchers must always keep working on a small dataset, seeing that the manual interpretation takes a long time (several weeks). However, they have to

investigate further the performance of the different feature extraction techniques, the exploration of unbalanced data methods, and other deep learning techniques, such as 1D-CNN and CNN-long short-term memory networks (CNN-LSTM).

6. Conclusions

This study proposes a hybrid VGG-16 and GBDT model for karst cavity segmentation. The study proposed a classifier that can identify cave facies by minimizing the error in their geometry based on the small dataset. The results revealed that LightGBM was the best model, with an F1-score ranging from 0.87 to 0.94 and a micro-G-Mean ranging from 0.92 to 0.96 compared to other models. Compared to others, the proposed model can learn and predict perfectly by preserving spatially consistent patterns and similar main features of reference images of any size and orientation. As a result, the F1-score, micro-G-Means, weight-G-Mean, and multi-class ROC-AUC curves show that the combination of these interactive approaches produces good results and can be used to distinguish different karst cavities and unconformities with a minimum value of 0.7. Based on CDF, LightGBM depicts favorable uncertainty in cave prediction.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/en16020643/s1 Figure S1: Predicted cave facies by the LightGBM model compared to references (a) and visualization of under- or overestimated cave area (b); Table S1: Selection ranges of each hyperparameter of GBDT models for caves and unconformities prediction; Table S2: F1-score of classifiers during training step; Table S3: Training times of models.

Author Contributions: Conceptualization, L.P., X.W. and H.H.; Methodology, A.K.F.K., X.W., A.K.M. and E.E.N.; Software, Z.W., F.J. and A.H.; Validation, X.W.; Formal analysis, A.K.M. and E.E.N.; Investigation, F.J., M.S. and A.H.; Data curation, A.K.F.K., L.P. and Z.W.; Writing – original draft, A.K.F.K.; Writing – review & editing, A.K.M., F.J., M.S., A.H., H.H. and E.E.N.; Visualization, M.S. and H.H.; Supervision, L.P.; Project administration, Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Authors do not have permission to share data.

Acknowledgments: We would like to thank China Petroleum & Chemical Corporation (SINOPEC) for providing samples, data, and permission to publish the results.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Firme, P.A.L.P.; Quevedo, R.J.; Roehl, D.; Pereira, L.C.; Cazarin, C.L. Mechanical behavior of carbonate reservoirs with single karst cavities. *Geomech. Energy Environ.* **2021**, *25*, 100209.
2. Garland, J.; Neilson, J.; Laubach, S.E.; Whidden, K.J. Advances in carbonate exploration and reservoir analysis. *Geol. Soc. Lond. Spec. Publ.* **2012**, *370*, 1–15. https://doi.org/10.1144/SP370.15.
3. Sha, F.; Xiao, L.; Mao, Z.; Jia, C. Petrophysical characterization and fractal analysis of carbonate reservoirs of the eastern margin of the pre-Caspian Basin. *Energies* **2019**, *12*, 78.
4. Wang, S.; Wang, S.; Yu, C.; Liu, H. Single Well Productivity Prediction Model for Fracture-Vuggy Reservoir Based on Selected Seismic Attributes. *Energies* **2021**, *14*, 4134.
5. Yang, L.I.; Zhijiang, K.; Zhaojie, X.U.E.; Zheng, S. Theories and practices of carbonate reservoirs development in China. *Pet. Explor. Dev.* **2018**, *45*, 712–722.
6. Zhiliang, H.; Shoutao, P.; Tao, Z. Controlling factors and genetic pattern of the Ordovician reservoirs in the Tahe area, Tarim Basin. *Oil Gas Geol.* **2010**, *31*, 743–752.
7. Xu, X.; Chen, Q.; Zhang, Y.; Wang, J.; Li, Y.; Kang, Z.; Wei, H.; Quan, L.; Zhang, Y. Research progress and prospect of Ordovician carbonate rocks in Tahe oilfield: Karst feature. *J. Pet. Explor. Prod. Technol.* **2021**, *11*, 3889–3902. https://doi.org/10.1007/s13202-021-01268-1.

8. Yanping, L.; Jing, L.; Xiangdong, X.; Guangxiao, D.; Yongli, L.; Cunge, L.; Zhenzhe, Z.; Yongqiang, H. Genetic mechanism of inner reservoirs of Yingshan Formation of Middle-Lower Ordovician in Tahe Oil Field, Tarim Basin. *Pet. Geol. Exp.* **2021**, *43*, 1031–1037.
9. Hu, X.; Li, Y.; Quan, L.; Kong, Q.; Wang, Y.; Lv, X.R. Three-dimensional geological modeling of factured-vuggy carbonate reservoirs: A case from the Ordovician reservoirs in Tahe-IV block, Tahe oilfield. *Oil Gas Geol.* **2013**, *34*, 383–387.
10. Yang, X.; Wang, X.; Tang, H.; Ding, Y.; Lv, H.; Liu, C. The early Hercynian paleo-karstification in the Block 12 of Tahe oilfield, northern Tarim Basin, China. *Carbonates Evaporites* **2014**, *29*, 251–261.
11. Yang, L.I.; Jiagen, H.O.U.; Yongqiang, L.I. Features and classified hierarchical modeling of carbonate fracture-cavity reservoirs. *Pet. Explor. Dev.* **2016**, *43*, 655–662.
12. Loucks, R.G. Paleocave carbonate reservoirs: Origins, burial-depth modifications, spatial complexity, and reservoir implications. *Am. Assoc. Pet. Geol. Bull.* **1999**, *83*, 1795–1834.
13. Zhiliang, H.E.; Jianfang, S.U.N.; Panhong, G.U.O.; Hehua, W.E.I.; Xinrui, L.Y.U.; Kelong, H.A.N. Construction of carbonate reservoir knowledge base and its application in fracture-cavity reservoir geological modeling. *Pet. Explor. Dev.* **2021**, *48*, 824–834.
14. He, J.; Li, A.; Wu, S.; Tang, R.; Lv, D.; Li, Y.; Li, X. Experimental investigation on injection and production pattern in fractured-vuggy carbonate reservoirs. *Energies* **2020**, *13*, 603.
15. He, Z.; Peng, S.; Zhang, T. Controls on reservoir formation in Ordovician of Tahe oilfield, Tarim basin, and combinational genetic mechanism. *Oil Gas Geol.* **2010**, *31*, 743–752.
16. Kuanzhi, Z.; Zhang, L.; Zheng, D.; Chonghao, S.; Qingning, D. A reserve calculation method for fracture-cavity carbonate reservoirs in Tarim Basin, NW China. *Pet. Explor. Dev.* **2015**, *42*, 277–282.
17. Tian, F.; Jin, Q.; Li, Y. A new logging recognition method of small fracture-cave and fills in fracture-cavity reservoirs in Tahe oilfield. *Oil Gas Geol.* **2012**, *33*, 900–908.
18. Sun, S.Z.; Zhou, X.; Yang, H.; Wang, Y.; Wang, D.; Liu, Z. Fractured reservoir modeling by discrete fracture network and seismic modeling in the Tarim Basin, China. *Pet. Sci.* **2011**, *8*, 433–445.
19. Pazzi, V.; Di Filippo, M.; Di Nezza, M.; Carlà, T.; Bardi, F.; Marini, F.; Fontanelli, K.; Intrieri, E.; Fanti, R. Integrated geophysical survey in a sinkhole-prone area: Microgravity, electrical resistivity tomographies, and seismic noise measurements to delimit its extension. *Eng. Geol.* **2018**, *243*, 282–293. https://doi.org/10.1016/j.enggeo.2018.07.016.
20. Torrese, P. Investigating karst aquifers: Using pseudo 3-D electrical resistivity tomography to identify major karst features. *J. Hydrol.* **2020**, *580*, 124257. https://doi.org/10.1016/j.jhydrol.2019.124257.
21. Zeid, N.A.; Bignardi, S.; Russo, P.; Peresani, M. Deep in a Paleolithic archive: Integrated geophysical investigations and laser-scanner reconstruction at Fumane Cave, Italy. *J. Archaeol. Sci. Rep.* **2019**, *27*, 101976. https://doi.org/10.1016/j.jasrep.2019.101976.
22. Martínez-Moreno, F.; Galindo-Zaldívar, J.; Baena, C.L.; González-Castillo, L.; Herrera, J.B.; Martínez-Martos, M.; Padial, Y.d.R.; Rodríguez, L.F.; Tendero-Salmerón, V.; Madarieta-Txurruka, A. Development and collapse of karstic cavities in folded marbles: Geomorphological and geophysical evidences in Nerja Cave (southern Spain). *J. Appl. Geophys.* **2021**, *187*, 104287. https://doi.org/10.1016/j.jappgeo.2021.104287.
23. Meng, Z.; Sun, Z.; Li, G. A case study of complex carbonate reservoir connectivity analysis, Tarim Basin, China. *Interpretation* **2021**, *9*, B77–B87.
24. Loule, J.-P.; Jifon, F.; Bioule, S.E.A.; Nguema, P.; Spofforth, D.; Carruthers, D.; Watkins, C.; Johnston, J. An opportunity to re-evaluate the petroleum potential of the Douala/Kribi-Campo Basin, Cameroon. *First Break* **2018**, *36*, 61–70. https://doi.org/10.3997/1365-2397.N0078.
25. Li, Z.; Wang, Y.; Yang, Z.; Li, H.; Yu, G. Identification of fractured carbonate vuggy reservoirs in the S48 well area using 3D 3C seismic technique: A case history from the Tarim Basin. *Geophysics* **2019**, *84*, B59–B74.
26. Tian, F.; Jin, Q.; Lu, X.; Lei, Y.; Zhang, L.; Zheng, S.; Zhang, H.; Rong, Y.; Liu, N. Multi-layered Ordovician paleokarst reservoir detection and spatial delineation: A case study in the Tahe Oilfield, Tarim Basin, Western China. *Mar. Pet. Geol.* **2016**, *69*, 53–73.
27. Li, X.; Chen, Q.; Wu, C.; Liu, H.; Fang, Y. Application of multi-seismic attributes analysis in the study of distributary channels. *Mar. Pet. Geol.* **2016**, *75*, 192–202. https://doi.org/10.1016/j.marpetgeo.2016.04.016.
28. Shan, X.; Tian, F.; Cheng, F.; Yang, C.; Xin, W. Spectral decomposition and a waveform cluster to characterize strongly heterogeneous paleokarst reservoirs in the Tarim Basin, China. *Water* **2019**, *11*, 256.
29. Xin, W.; Tian, F.; Shan, X.; Zhou, Y.; Rong, H.; Yang, C. Application of geologically constrained machine learning method in characterizing paleokarst reservoirs of tarim basin, China. *Water* **2020**, *12*, 1765.
30. Méndez, J.N.; Jin, Q.; Zhang, X.; González, M.; Kashif, M.; Boateng, C.D.; Zambrano, M. Rock type prediction and 3D modeling of clastic paleokarst fillings in deeply-buried carbonates using the Democratic Neural Networks Association technique. *Mar. Pet. Geol.* **2021**, *127*, 104987.
31. Wu, X.; Yan, S.; Qi, J.; Zeng, H. Deep learning for characterizing paleokarst collapse features in 3-D seismic images. *J. Geophys. Res. Solid Earth* **2020**, *125*, e2020JB019685.
32. Zhang, G.; Lin, C.; Ren, L.; Li, S.; Cui, S.; Wang, K.; Sun, Y. Seismic characterization of deeply buried paleocaves based on Bayesian deep learning. *J. Nat. Gas. Sci. Eng.* **2022**, *97*, 104340.

33. Wu, S.; Wang, Q.; Zeng, Q.; Zhang, Y.; Shao, Y.; Deng, F.; Liu, Y.; Wei, W. Automatic extraction of outcrop cavity based on a multiscale regional convolution neural network. *Comput. Geosci.* **2022**, *160*, 105038.
34. Tang, J.; Fan, B.; Xu, G.; Xiao, L.; Tian, S.; Luo, S.; Weitz, D. A new tool for searching sweet spots by using gradient boosting decision trees and generative adversarial networks. In Proceedings of the International Petroleum Technology Conference, Dhahran, Saudi Arabia, 13 January 2020.
35. Asante-Okyere, S.; Shen, C.; Ziggah, Y.Y.; Rulegeya, M.M.; Zhu, X. A novel hybrid technique of integrating gradient-boosted machine and clustering algorithms for lithology classification. *Nat. Resour. Res.* **2020**, *29*, 2257–2273.
36. Sun, J.; Chen, M.; Li, Q.; Ren, L.; Dou, M.; Zhang, J. A new method for predicting formation lithology while drilling at horizontal well bit. *J. Pet. Sci. Eng.* **2021**, *196*, 107955.
37. Ruiyi, H.A.N.; Zhuwen, W.; Wenhua, W.; Fanghui, X.U.; Xinghua, Q.I.; Yitong, C.U.I. Lithology identification of igneous rocks based on XGboost and conventional logging curves, a case study of the eastern depression of Liaohe Basin. *J. Appl. Geophys.* **2021**, *195*, 104480.
38. Liu, J.-J.; Liu, J.-C. Integrating deep learning and logging data analytics for lithofacies classification and 3D modeling of tight sandstone reservoirs. *Geosci. Front.* **2022**, *13*, 101311.
39. Liu, J.-J.; Liu, J.-C. An intelligent approach for reservoir quality evaluation in tight sandstone reservoir using gradient boosting decision tree algorithm-A case study of the Yanchang Formation, mid-eastern Ordos Basin, China. *Mar. Pet. Geol.* **2021**, *126*, 104939.
40. Gu, J.; Liu, W.; Zhang, K.; Zhai, L.; Zhang, Y.; Chen, F. Reservoir production optimization based on surrograte model and differential evolution algorithm. *J. Pet. Sci. Eng.* **2021**, *205*, 108879.
41. Liu, W.; Liu, W.D.; Gu, J. Predictive model for water absorption in sublayers using a Joint Distribution Adaption based XGBoost transfer learning method. *J. Pet. Sci. Eng.* **2020**, *188*, 106937.
42. Otchere, D.A.; Ganat, T.O.A.; Gholami, R.; Lawal, M. A novel custom ensemble learning model for an improved reservoir permeability and water saturation prediction. *J. Nat. Gas. Sci. Eng.* **2021**, *91*, 103962.
43. Pan, S.; Zheng, Z.; Guo, Z.; Luo, H. An optimized XGBoost method for predicting reservoir porosity using petrophysical logs. *J. Pet. Sci. Eng.* **2022**, *208*, 109520.
44. Morozov, A.D.; Popkov, D.O.; Duplyakov, V.M.; Mutalova, R.F.; Osiptsov, A.A.; Vainshtein, A.L.; Burnaev, E.V.; Shel, E.V.; Paderin, G.V. Data-driven model for hydraulic fracturing design optimization: Focus on building digital database and production forecast. *J. Pet. Sci. Eng.* **2020**, *194*, 107504.
45. Tang, J.; Fan, B.; Xiao, L.; Tian, S.; Zhang, F.; Zhang, L.; Weitz, D. A New Ensemble Machine-Learning Framework for Searching Sweet Spots in Shale Reservoirs. *SPE J.* **2021**, *26*, 482–497.
46. Zhong, R.; Johnson, R., Jr.; Chen, Z. Generating pseudo density log from drilling and logging-while-drilling data using extreme gradient boosting (XGBoost). *Int. J. Coal Geol.* **2020**, *220*, 103416.
47. Pirizadeh, M.; Alemohammad, N.; Manthouri, M.; Pirizadeh, M. A new machine learning ensemble model for class imbalance problem of screening enhanced oil recovery methods. *J. Pet. Sci. Eng.* **2021**, *198*, 108214.
48. Qaid, T.S.; Mazaar, H.; Al-Shamri, M.Y.H.; Alqahtani, M.S.; Raweh, A.A.; Alakwaa, W. Hybrid Deep-Learning and Machine-Learning Models for Predicting COVID-19. *Comput. Intell. Neurosci.* **2021**, *2021*, 9996737. https://doi.org/10.1155/2021/9996737.
49. Ruan, Z.; Yu, B.; Wang, L.; Pan, Y.; Tan, G. Prediction of buried calcite dissolution in the Ordovician carbonate reservoir of the Tahe Oilfield, NW china: Evidence from formation water. *Geochemistry* **2013**, *73*, 469–479. https://doi.org/10.1016/j.chemer.2013.03.004.
50. Chen, Q.; Zhao, Y.; Li, G.; Chu, C.; Wang, B. Features and controlling factors of epigenic karstification of the Ordovician carbonates in Akekule Arch, Tarim Basin. *J. Earth Sci.* **2012**, *23*, 506–515.
51. Tian, F.; Wang, Z.; Cheng, F.; Xin, W.; Fayemi, O.; Zhang, W.; Shan, X. Three-dimensional geophysical characterization of deeply buried paleokarst system in the Tahe Oilfield, Tarim Basin, China. *Water* **2019**, *11*, 1045.
52. Ding, Z.; Wang, R.; Chen, F.; Yang, J.; Zhu, Z.; Yang, Z.; Sun, X.; Xian, B.; Li, E.; Shi, T.; et al. Origin, hydrocarbon accumulation and oil-gas enrichment of fault-karst carbonate reservoirs: A case study of Ordovician carbonate reservoirs in South Tahe area of Halahatang oilfield, Tarim Basin. *Pet. Explor. Dev.* **2020**, *47*, 306–317. https://doi.org/10.11698/PED.2020.02.07.
53. Wu, J.; Fan, T.; Gomez-Rivas, E.; Gao, Z.; Yao, S.; Li, W.; Zhang, C.; Sun, Q.; Gu, Y.; Xiang, M. Impact of pore structure and fractal characteristics on the sealing capacity of Ordovician carbonate cap rock in the Tarim Basin, China. *Mar. Pet. Geol.* **2019**, *102*, 557–579.
54. Zhang, S.; Qiang, J.I.N.; Jianfang, S.U.N.; Hehua, W.E.I.; Cheng, F.; Zhang, X. Formation of hoodoo-upland on Ordovician karst slope and its significance in petroleum geology in Tahe area, Tarim Basin, NW China. *Pet. Explor. Dev.* **2021**, *48*, 354–366.
55. Rao, H.; Shi, X.; Rodrigue, A.K.; Feng, J.; Xia, Y.; Elhoseny, M.; Yuan, X.; Gu, L. Feature selection based on artificial bee colony and gradient boosting decision tree. *Appl. Soft Comput. J.* **2019**, *74*, 634–642. https://doi.org/10.1016/j.asoc.2018.10.036.
56. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232.
57. Li, D.; Faradonbeh, R.S.; Lv, A.; Wang, X.; Roshan, H. A data-driven field-scale approach to estimate the permeability of fractured rocks. *Int. J. Min. Reclam. Environ.* **2022**, *36*, 671–687. https://doi.org/10.1080/17480930.2022.2086769.
58. Zhu, X.; Chu, J.; Wang, K.; Wu, S.; Yan, W.; Chiam, K. Prediction of rockhead using a hybrid N-XGBoost machine learning framework. *J. Rock Mech. Geotech. Eng.* **2021**, *13*, 1231–1245.

59. Cui, J.-F.; Xia, H.; Zhang, R.; Hu, B.-X.; Cheng, X.-G. Optimization scheme for intrusion detection scheme GBDT in edge computing center. *Comput. Commun.* **2020**, *168*, 136–145. https://doi.org/10.1016/j.comcom.2020.12.007.
60. Ma, X.; Sha, J.; Wang, D.; Yu, Y.; Yang, Q.; Niu, X. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electron. Commer. Res. Appl.* **2018**, *31*, 24–39.
61. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural. Inf. Process. Syst.* **2017**, *30*, 3146–3154.
62. Koponen, J.-P. Predicting Lead Times of Purchase Orders Using Gradient Boosting Machine. Master's Thesis, Lappeenranta-Lahti University of Technology LUT, Lappeenranta, Finland, 2020.
63. González, S.; García, S.; del Ser, J.; Rokach, L.; Herrera, F. A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Inf. Fusion* **2020**, *64*, 205–237.
64. Wang, M.; Yue, L.; Yang, X.; Wang, X.; Han, Y.; Yu, B. Fertility-LightGBM: A fertility-related protein prediction model by multi-information fusion and light gradient boosting machine. *Biomed. Signal Process. Control* **2021**, *68*, 102630.
65. Liu, J.; Gao, Y.; Hu, F. A fast network intrusion detection system using adaptive synthetic oversampling and LightGBM. *Comput. Secur.* **2021**, *106*, 102289.
66. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. *arXiv* **2017**, arXiv:1706.09516.
67. Dorogush, A.V.; Ershov, V.; Gulin, A. CatBoost: Gradient boosting with categorical features support. *arXiv* **2018**, arXiv:1810.11363.
68. Dhananjay, B.; Sivaraman, J. Analysis and classification of heart rate using CatBoost feature ranking model. *Biomed. Signal Process. Control* **2021**, *68*, 102610.
69. Rahman, S.; Irfan, M.; Raza, M.; Ghori, K.M.; Yaqoob, S.; Awais, M. Performance analysis of boosting classifiers in recognizing activities of daily living. *Int. J. Environ. Res. Public Health* **2020**, *17*, 1082.
70. Huang, G.; Wu, L.; Ma, X.; Zhang, W.; Fan, J.; Yu, X.; Zeng, W.; Zhou, H. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. *J. Hydrol.* **2019**, *574*, 1029–1041.
71. Geldmacher, J.E. Convolutional Neural Networks for Feature Extraction and Automated Target Recognition in Synthetic Aperture Radar Images. Master's Thesis, Naval Postgraduate School, Monterey, CA, USA, 2020.
72. Murali, S.; Deepu, R.; Shivamurthy, R.C. ResNet-50 vs VGG-19 vs Training from Scratch: A comparative analysis of the segmentation and classification of Pneumonia from chest x-ray images. *Glob. Transit. Proc.* **2021**, *2*, 375–381.
73. Rahman, M.; Cao, Y.; Sun, X.; Li, B.; Hao, Y. Deep pre-trained networks as a feature extractor with XGBoost to detect tuberculosis from chest X-ray. *Comput. Electr. Eng.* **2021**, *93*, 107252.
74. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
75. Grycza, J.; Horna, D.; Klimczak, H.; Lango, M.; Pluciński, K.; Stefanowski, J. Multi-Imbalance: Open Source Python Toolbox for Multi-class Imbalanced Classification. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Spring: New York, NY, USA, 2021; Volume 12461. https://doi.org/10.1007/978-3-030-67670-4_36.
76. Rodríguez, J.J.; Diez-Pastor, J.-F.; Arnaiz-Gonzalez, A.; Kuncheva, L.I. Random Balance ensembles for multiclass imbalance learning. *Knowl. Based Syst.* **2020**, *193*, 105434.
77. Chen, X.; Yu, D.; Fan, X.; Wang, L.; Chen, J. Multiclass Classification for Self-Admitted Technical Debt Based on XGBoost. *IEEE Trans. Reliab.* **2021**, *71*, 1309–1324.
78. Berrar, D. Performance Measures for Binary Classification. In *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*; Elsevier: Amsterdam, The Netherlands, 2019; Volume 1–3, pp. 546–560. https://doi.org/10.1016/B978-0-12-809633-8.20351-8.
79. Bradley, A.P. The Use of the Area under the Roc Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognit.* **1997**, *30*, 1145–1159.
80. Hou, S.; Liu, Y.; Yang, Q. Real-time prediction of rock mass classification based on TBM operation big data and stacking technique of ensemble learning. *J. Rock Mech. Geotech. Eng.* **2021**, *14*, 123–143.
81. Zhou, J.; Li, E.; Yang, S.; Wang, M.; Shi, X.; Yao, S.; Mitri, H.S. Slope stability prediction for circular mode failure using gradient boosting machine approach based on an updated database of case histories. *Saf. Sci.* **2019**, *118*. https://doi.org/10.1016/j.ssci.2019.05.046.
82. Jain, S.; Saha, A. Improving performance with hybrid feature selection and ensemble machine learning techniques for code smell detection. *Sci. Comput. Program* **2021**, *212*, 102713.
83. El Alani, O.; Abraim, M.; Ghennioui, H.; Ghennioui, A.; Ikenbi, I.; Dahr, F.-E. Short term solar irradiance forecasting using sky images based on a hybrid CNN–MLP model. *Energy Rep.* **2021**, *7*, 888–900.
84. Del Frate, F.; Pacifici, F.; Schiavon, G.; Solimini, C. Use of neural networks for automatic classification from high-resolution images. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 800–809.

85. Haghighat, F. Predicting the trend of indicators related to COVID-19 using the combined MLP-MC model. *Chaos Solitons Fractals* **2021**, *152*, 111399.
86. Zheng, S.; Yang, M.; Kang, Z.; Liu, Z.; Long, X.; Liu, K.; Li, X.; Zhang, S. Controlling factors of remaining oil distribution after water flooding and enhanced oil recovery methods for fracture-cavity carbonate reservoirs in Tahe Oilfield. *Pet. Explor. Dev.* **2019**, *46*, 786–795.