

Subsurface Natural Fracture Identification Using an Integrated Ensemble Learning Method

Guoqing Lu¹, Lianbo Zeng, Guoping Liu, Xiaoxuan Chen, Mehdi Ostadhassan², Xiaoyu Du, and Yangkang Chen³, *Senior Member, IEEE*

Abstract—Natural fractures play a crucial role in the storage and seepage of oil shale. However, identifying fractures using conventional logging techniques presents challenges due to complex response characteristics and severe data imbalance. Here, we propose a highly accurate integrated ensemble learning method, called BSI-extreme gradient boosting (XGBoost), for identifying the natural fracture development, which combines several steps including isolation forests (iForests), synthetic minority oversampling techniques (SMOTEs), and XGBoost, and incorporates rock brittleness as a controlling factor in the model construction process. The proposed model effectively addresses several challenges encountered in fracture identification, including complex logging response characteristics, low precision and recall of fractured labels, and excessive sensitivity of ensemble learning to noise. To do so, the relationship between fracture density and brittle mineral content is analyzed through core analysis and X-ray diffraction (XRD). Then, conventional logging and rock brittleness are used as features for training the model. Herein, by screening the outliers of iForest, SMOTE oversampling, and feature selection, optimal hyperparameters of the model are obtained through the grid search method. The results demonstrated that using BSI-XGBoost, the testing set achieved an accuracy of 92.45%. Comparatively, this accuracy is 4.86% higher than the original XGBoost model and 3.73% higher than the B-XGBoost model, which incorporated brittleness curves but did not include oversampling and outlier removal. Collectively, this workflow provided an effective method for intelligent identification of fractures in oil shale with high accuracy based on easily accessible conventional logging curves.

Index Terms—BSI-extreme gradient boosting (XGBoost), fracture identification, geoscience, image processing, integrated ensemble learning.

Received 29 October 2024; revised 14 November 2024; accepted 30 December 2024. Date of publication 3 January 2025; date of current version 16 January 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 42090025 and Grant 42302148 and in part by the Sponsored by China National Petroleum Corporation (CNPC) Innovation Found under Grant 2023DQ02-0103. (Corresponding author: Lianbo Zeng.)

Guoqing Lu is with the College of Geoscience, China University of Petroleum, Beijing 102249, China, and also with the Civil Engineering Department, Federal University of Rio de Janeiro, Rio de Janeiro 21941-853, Brazil (e-mail: zsdhldgq@126.com).

Lianbo Zeng and Xiaoyu Du are with the College of Geoscience, China University of Petroleum, Beijing 102249, China (e-mail: lbzeng@sina.com; duxiaoyu2016@163.com).

Guoping Liu is with the School of Energy Resources, China University of Geosciences, Beijing 100083, China (e-mail: liugp1228@sina.com).

Xiaoxuan Chen is with PetroChina Xinjiang Oilfield Company, Kalamay, Xinjiang 834000, China (e-mail: cxxsxs@petrochina.com.cn).

Mehdi Ostadhassan is with the Institute of Geosciences, Marine and Land Geomechanics and Geotectonics, Christian-Albrechts-Universität, 24118 Kiel, Germany (e-mail: mehdi.ostadhassan@nepu.edu.cn).

Yangkang Chen is with the Bureau of Economic Geology, The University of Texas at Austin, Austin, TX 78713 USA (e-mail: chenyk2016@gmail.com).

Digital Object Identifier 10.1109/TGRS.2024.3525191

I. INTRODUCTION

WITH recent increase in oil and gas exploration and development of unconventional resources, shale oil has become an important supplier of the global energy need [1], [2], [3], [4]. To date, main shale oil exploration and operational activities have taken place in marine shale. However, China is rich in continental shale oil as well, with the resulting liquid petroleum characterized by high viscosity and poor mobility [5], [6]. Due to oil shale's low porosity and permeability, effective flow paths system solely based on matrix pore network is highly restricted. Natural fractures refer to the discontinuities that naturally exist in reservoir rocks due to tectonic deformation or physical diagenesis. Open natural fractures significantly contribute to the enrichment, migration, and preservation of oil and gas in continental shale with low matrix porosity and permeability [7], [8]. Therefore, the study of natural fractures would be necessary for more successful exploration and development of continental oil shale not only in China but around the globe. In addition, lithofacies change abruptly in oil shale which introduces a high degree of heterogeneity, which necessitates the studies of natural fractures and their delineation.

Core and image logs are the two main sources of data for the identification and interpretation of fractures. However, the cost of coring as well as image logging limits the availability of such data [9]. This eventually will challenge to systematically delineate and study the target interval with high accuracy regarding its petrophysical properties, specifically the presence of natural fractures. On the contrary, conventional logging is commonly deployed in wells, offering high vertical resolution and continuity, and providing a wealth of data from various angles, making it a suitable and easily accessible database for fracture interpretation [10]. In addition, other factors such as the pore fluids and clay content also affect the logging response, resulting in a considerable overlap between the fractured and nonfractured intervals, which makes identification of fractures from such source of data a complex and nonlinear problem [10], [11], [12], [13], [14].

To overcome these obstacles, machine learning approaches have demonstrated the ability to delineate and model the nonlinear relationship between fractures and logging curves [10], [15], [16], [17], [18], [19], [20]. Common machine learning methods can be divided into single classifier and ensemble classifier methods [21], [22], [23]. The single classifier methods that have been used for fracture identification include

the Bayesian network (BN), K -nearest neighbors (KNNs), support vector machine (SVM), etc. BN is a probability graph model that represents a set of variables and their conditional dependencies through a directed acyclic graph. This algorithm is highly suitable for recording events and predicting the likelihood of several known causes that can contribute to a factor [24]. Based on this feature, the BN has been utilized to identify the fractures in the Appalachian Basin, to reveal the correlation and causal relationship between the input logging curves and the fractures [25]. KNNs is a method of calculating the distance between a new sample and a known sample point, selecting K closest samples, and counting the majority of the sample labels [26]. For example, KNN was used to identify fractures in Changxing and Feixianguan Formation of Puguang Gas Field, where the results were consistent with the observed fractures in the core [27]. SVM determines the samples that are near the decision boundary as support vectors in high-dimensional space [28] which has been used in fracture identification. In the application of this method, the maximum margin hyperplane was solved by the logging data that was calibrated by core inspections as the boundary of fracture identification, which was determined by the support vector [29], [30].

Reducing the uncertainty of model classification is the key to applying artificial intelligence methods for fracture identification. In most cases, the performance of ensemble learning classifiers is superior to single classifiers, which are better suited for solving nonlinear problems [31], [32], [33], [34]. The ensemble learning methods combine a set of weak classifiers (called individual classifiers or subclassifiers) into a strong classifier, which jointly determines the classification results through a joint decision mechanism, and overcomes the problems of overfitting and low classification accuracy. [35], [36]. Two commonly used methods in ensemble learning are Bagging [37] and Boosting [38]. Bagging is a random sampling with placement, which keeps the dataset size unchanged during the sampling process. In the training process, each weak learner has the same weight and is independent of each other. The final classification results can be obtained by averaging or voting mechanism [35], [37]. The representative Bagging method is random forest (RF). In the application of fracture identification in carbonate reservoirs in the Appalachian Basin of North America, using limited conventional logging data, RF successfully identified fractures with an accuracy of about 5%–9% higher than BN [25]. The boosting algorithm generates weak learners through a sequence approach, where the later ones are related to the previous one. The model first uses the initial weights to train a weak classifier from the training data, and then updates the weights of the samples based on the learning error rate of the weak classifiers [39]. Samples with higher error rates will receive higher weights and attract more attention in weak classifiers. Every time a weak classifier is added, the weight is readjusted and looped until a specified number of classifiers are obtained [40]. Then, by combining strategies for integration, the final strong classifier can be obtained. Notably, extreme gradient boosting (XGBoost) is one of the most widely used boosting methods due to its generalized effectiveness.

Directly applying the original XGBoost to fracture identification tasks can lead to various challenges as follows.

- 1) The presence of noisy points, like anomalies, can complicate the error correction process of XGBoost during training. These anomaly points may be misclassified, resulting in decreased classification performance. Furthermore, as the model iteratively improves, it amplifies the impact of these errors, further hindering its ability to accurately classify data points.
- 2) The data used for fracture identification is imbalanced, with a significantly larger number of nonfracture samples compared to fracture samples.

Training a model directly on this imbalanced dataset can result in poor identification performance for fractured samples. To address the aforementioned challenges, we propose a novel integrated ensemble learning method, called the BSI-XGBoost method. This method incorporates rock brittleness, a key factor in fracture development, as a characteristic of the training data. Additionally, it integrates isolation forests (iForests), synthetic minority over-sampling technique (SMOTE), and XGBoost into the construction process of the identification model, offering several advantages as follows.

- 1) The BSI-XGBoost integrates geological understanding into model construction by utilizing rock brittleness as a feature in training. Considering rock brittleness as one of the controlling factors in fracture development enhances the model's ability to accurately classify fractures.
- 2) By including the SMOTE algorithm, the BSI-XGBoost method overcomes the imbalanced data problem caused by a smaller number of fracture samples compared to nonfracture samples. This improves the precision and recall rates for identifying fracture samples.
- 3) The BSI-XGBoost effectively addresses the issue of error amplification during iterative boosting by introducing the iForest algorithm and feature selection. This ensures that the model maintains high classification performance, avoiding a decrease in accuracy.

The objective of this study is the oil shale of the Fengcheng Formation in the Mahu Sag of the Junggar Basin, western China. To achieve a unique solution for the fracture identification problem, it is necessary to first process and optimize conventional logging data. Then, we should establish an intelligent identification model based on our proposed BSI-XGBoost method, by using core and image logs to label the conventional logging data. This process will enable us to achieve longitudinal identification of fracture development in the target formation in the study area.

II. METHODOLOGY

A. Basic Idea of the Fracture Identification Method

The dataset employed for training the model in this study comprises well-logging curves, interpretation curves, and lithological data. Each curve is constituted by numerous logging sampling points spaced at 0.125 m intervals. Label data are generated through core observations and image log interpretations. Essentially, the fracture identification model operates as a multiple classifier, categorizing each logging

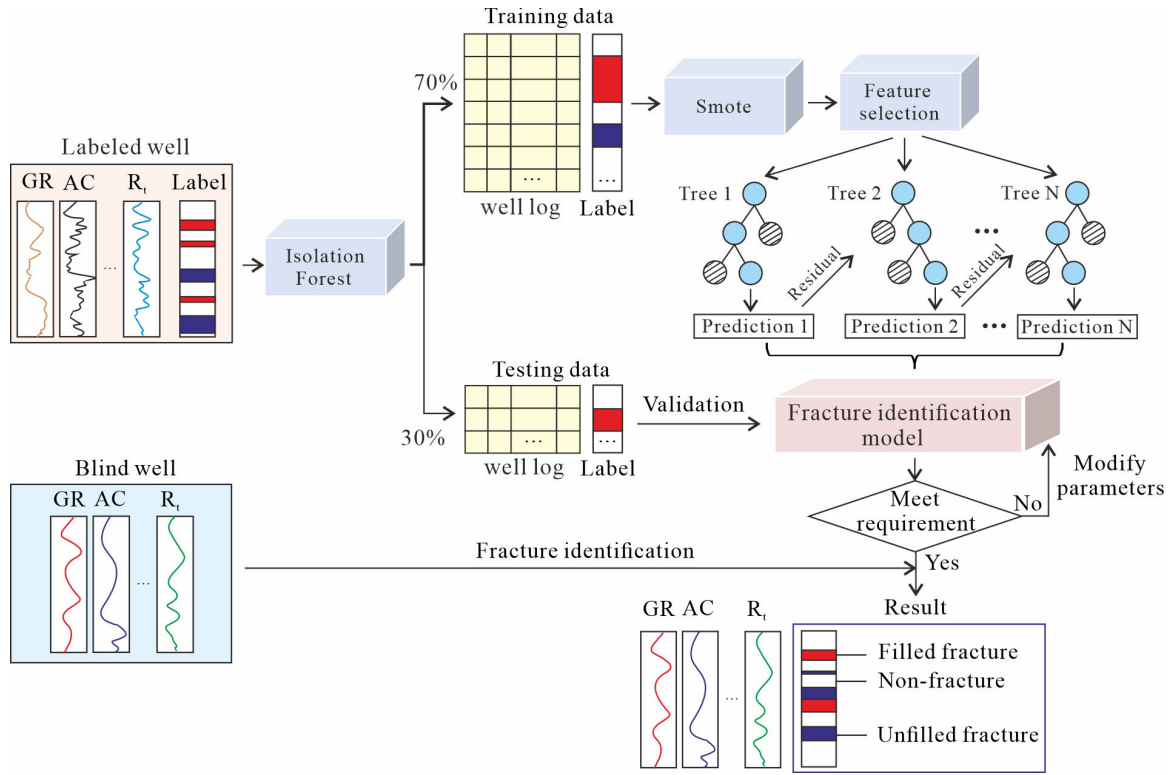


Fig. 1. Process of the BSI-XGBoost for fracture identification, which includes data preparation, noise suppression, and decision tree ensemble classification.

sampling point as either a filled fracture, unfilled fracture, or nonfracture. The general workflow of the intelligent fracture identification discussed in this article, as shown in Fig. 1, can be summarized as three steps.

1) *Data Preparation*: In this study, core samples, image logs, and conventional logs are collected from the key wells of the Fengcheng Formation in Mahu Sag, Junggar Basin, western China, for fracture identification. The rock brittleness data and conventional logs will be used as the training and validation datasets for the identification model. The labels for the dataset were derived from core observations and the interpretation of image logging results related to fractures. We use the min-max normalization method to process nine curves, including Rock Brittleness (Brittleness), gamma ray (GR), caliper (CAL), sonic log (AC), density log (DEN), compensated neutron log (CNL), invasion zone formation resistivity (R_i), true formation resistivity (R_t), and flushed zone formation resistivity (R_{xo}). The screened dataset is randomly divided into a training set and a testing set at 7:3. In addition, we select another well that is not involved in model construction as a test well.

2) *Noise Suppression, Oversampling, and Feature Selection*: Due to the imbalanced nature of the dataset, where the number of fractures is significantly smaller than nonfractures, the SMOTE algorithm is employed to balance the distribution of samples with different labels in the training set through oversampling. To mitigate the XGBoost method's sensitivity to noise, two strategies are implemented. First, the iForest method is utilized to identify and eliminate abnormal data points from the standardized dataset, ensuring their minimal impact on the model's performance. Second, the BSI-XGBoost

model employs feature selection techniques on multiple logging curves to identify the most influential features crucial for accurate fracture identification. By reducing the number of features, the model is less susceptible to noise. Eliminating less significant features effectively helps noise elimination and enhances the model's performance.

3) *Decision Tree Ensemble Classification*: The method includes multiple decision trees, where we use the co-determination strategy to obtain the final classification result, and the initial training model is obtained after multiple iterations. Herein, the remaining 30% of the data are used to examine the classification performance of the model as the test data. If the identification accuracy of the model for the testing data is too low, it would be necessary to tune the parameters again. Ultimately when the expected identification accuracy is achieved, the model is output and used for test well fracture identification. The criterion that is used to measure the accuracy of the model is explained later in the text.

B. Algorithm Details

1) *Isolation Forest*: iForest is an algorithm for data anomaly detection using binary trees [41]. In essence, the algorithm is that anomalous data points are easier to separate from the rest of the sample [41]. A set, denoted as $X = \{x_1, \dots, x_n\}$, is created for well B1 logging sampling points. Each sampling point is represented as x_1, \dots, x_n , forming a 9-D vector that includes a rock brittleness curve and eight conventional well logging curves. The set X of logging sampling points undergoes subsampling, resulting in t subsets X' . Multiple isolation trees (iTrees) are then constructed based on these

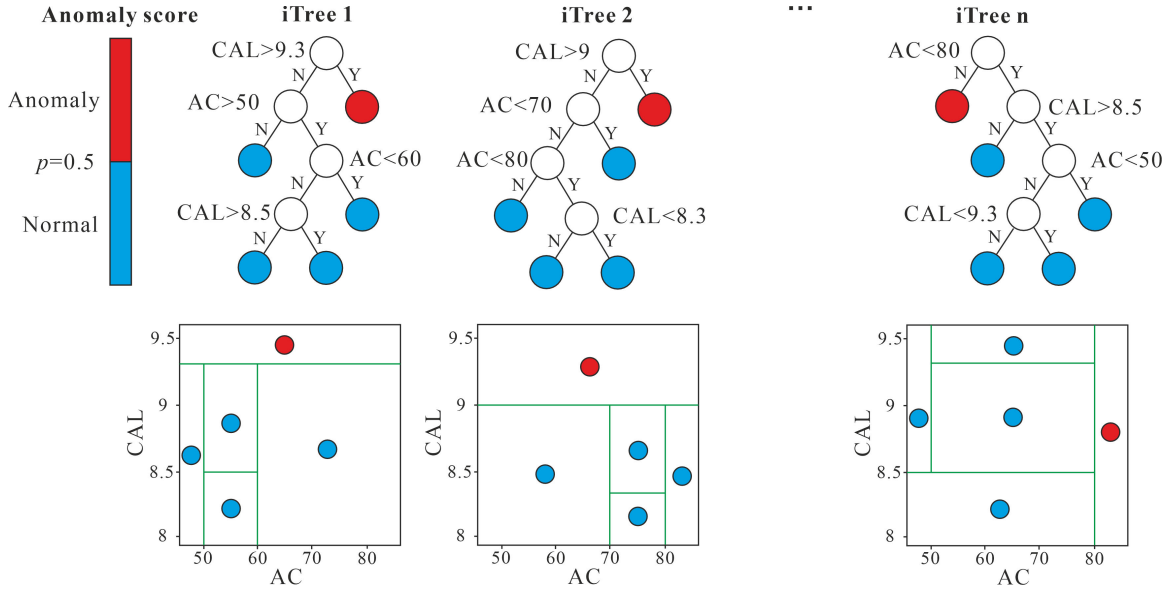


Fig. 2. Growth process of iTrees can be illustrated using the example of a 2-D feature matrix composed of CAL and ac .

subsets. As shown in Fig. 2, “iTrees” are data structures characterized by the following properties.

First, to isolate a data point, the algorithm recursively generates partitions on the sample by randomly selecting a feature (rock brittleness or conventional well log) in the dataset and then randomly selecting a split value p between the minimum and maximum values allowed for that feature. The determination of child nodes (T_l and T_r) is based on comparing the feature attribute values q and p of the node T (logging sampling point) within the tree. The algorithm recursively partitions X' by randomly selecting a q and a split value p until either the node contains only one instance, or all data in the node have identical values.

Second, when the iTree is fully grown, each node in X is isolated at one of the external nodes. Intuitively, the anomalous points are those that are easier to isolate, and thus have a smaller path length in the tree. The path length $h(x)$ of point $x \in X$ is defined as the number of edges (x) traverses from the root node to reach an external node. $h(x)$ can be expressed as

$$h(x) = e + c(T.size) \quad (1)$$

where e represents the number of edges traversed by sample x from the root node to a leaf node in the tree. $T.size$ denotes the number of samples that share the same leaf node as sample x . $c(T.size)$ can be seen as a correction value that indicates the average path length for constructing a binary tree with $T.size$ samples. The calculation formula for $c(n)$ is as follows:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \quad (2)$$

where $H(i)$ is the harmonic number and it can be estimated by $\ln(i) + 0.5772156649$ (Euler's constant).

As $c(n)$ is the average of $h(x)$ given n , we use it to normalize $h(x)$. The anomaly score s of an instance x is defined as

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}} \quad (3)$$

where $E(h(x))$ is the average of $h(x)$ from a collection of iTrees.

2) *SMOTE*: SMOTE is an over-sampling approach in which the minority class is over-sampled by creating “synthetic” examples rather than by over-sampling with replacement [42], which is also known as data augmentation for expanding the training database [43], [44]. SMOTE is used to oversample the minority samples in the training set. As shown in Fig. 3, the process is as follows: first, take a sample from the dataset, and consider its KNNs (in feature space). Second, take the vector between one of those k neighbors, and the current data point. Multiply this vector by a random number x which lies between 0, and 1. Add this to the current data point to create the new, synthetic data point [45].

3) *XGBoost*: XGBoost [46], [47] is a common boosting algorithm combining multiple weak classifiers to form a strong classifier. The following principle was used for fracture identification in this article.

The training samples consist of conventional well logs, rock brittleness, and labels, which can be represented by $D = \{X_s, Y\}$. In this representation, X_s represents the features that have undergone outlier detection and oversampling. $X_s = (x_1, x_2, \dots, x_m)^T$ represents the sample dataset containing m samples, where the feature vector of the i th sample, denoted by x_i , contains nine features. $Y = (y_1, y_2, \dots, y_m)^T$ represents the label matrix, where y_i represents the label corresponding to the x_i of the i th sample.

The prediction model after t iterations is represented as

$$\hat{y}_i^{(t)} = \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i). \quad (4)$$

In the above formula, $f_k(x_i)$ represents a CART tree, $\hat{y}_i^{(t-1)}$ is the prediction model after $t-1$ iterations.

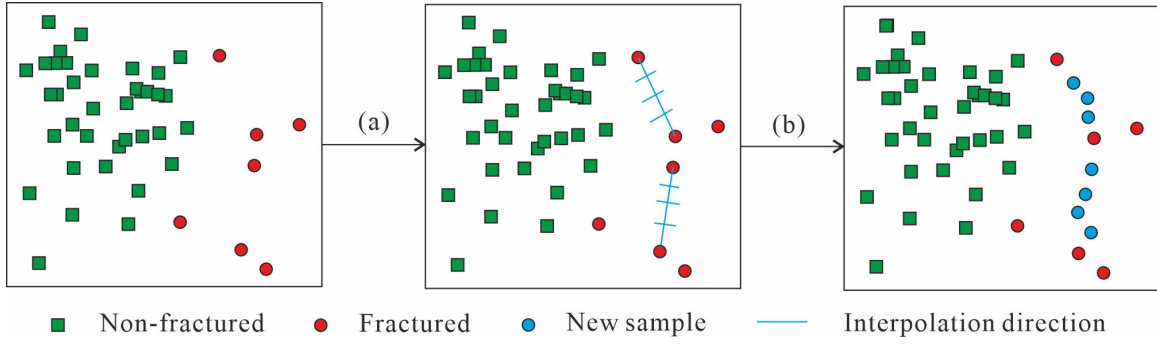


Fig. 3. SMOTE oversampling process. (a) Take the vector between one of those k neighbors, and the current data point. (b) Add this to the current data point to create the new, synthetic data point.

The corresponding objective function of the prediction model after t iterations

$$\tilde{L}^{(t)} = \sum_{i=1}^N \left(g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right) + \Omega(f_t) \quad (5)$$

$$g_i = \frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}, \quad h_i = \frac{\partial^2 L(y_i, \hat{y}_i^{(t-1)})}{\partial (\hat{y}_i^{(t-1)})^2} \quad (6)$$

$\Omega(\cdot)$ is the regularization term. The purpose of introducing regularization terms is to prevent overfitting. $L(\cdot)$ is the selected loss function.

The objective function of the prediction model is represented by leaf nodes

$$\tilde{L}^{(t)} = \sum_{j=1}^J \left[G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right] + \gamma J \quad (7)$$

$$G_j = \sum_{i \in I_j} g_i, \quad H_j = \sum_{i \in I_j} h_i \quad (8)$$

w_j represents the score of leaf node j , J represents the total number of leaf nodes, w represents the weight of the leaf node, γ represents the leaf node coefficient, and I_j represents the set of sample indices on leaf node j .

XGBoost defines the indicators for feature selection and segmentation point selection. For a specific splitting scheme, the gain obtained can be calculated by the formula

$$\text{Gain} = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma. \quad (9)$$

In the equation, $(G_L^2/H_L + \lambda)$, $(G_R^2/H_R + \lambda)$, $((G_L + G_R)^2/H_L + H_R + \lambda)$ represent the scores obtained from left subtree, right subtree, and undivided, respectively, while γ represents the cost of adding new leaf nodes. When $\text{Gain} < 0$, the node should not be divided into left and right branches. XGBoost uses the value calculated by this equation as the splitting condition and selects the most suitable split with the largest change.

C. Performance Evaluation Indicators

Based on model optimization, it is necessary to establish a unified evaluation index for fracture identification. The model can be quantitatively evaluated by using accuracy

TABLE I
DEFINITIONS OF PERFORMANCE EVALUATION INDICATORS

Evaluation indicators	Mathematical expression
Accuracy	$A_c = (TP + TN)/(TP + TN + FP + FN)$
Precision	$P_r = TP/(TP + FP)$
Recall	$R_e = TP/(TP + FN)$

* Note: TP = True Positives; FP = False Positives; TN = True Negatives; FN = False Negatives. TP , FP , TN , and FN represent the numbers of positive samples correctly predicted, negative samples incorrectly predicted as positive samples, positive samples incorrectly predicted as negative samples and negative samples correctly predicted, respectively.

(A_c), precision (P_r), and recall (R_e). As shown in Table I, accuracy (A_c) is the ratio of correctly classified samples to the total samples, reflecting the comprehensive classification performance of all samples in the model [48]. Precision (P_r) refers to the ratio of correctly classified samples to samples classified into that class, reflecting the accuracy of specific classification results [49]. The recall (R_e) is the ratio of correctly classified samples to the total sample in that class, reflecting the classification accuracy of a specific class [49].

D. Workflow of Fracture Identification

The proposed BSI-XGBoost method finds the existing non-linear relationship between fractures and conventional logging curve responses, to improve the accuracy of natural fracture identification from such highly accessible and available data in the petroleum industry. The workflow is shown in Fig. 4. The input data include core, image logs, and conventional log data. Due to the significant variation in the conventional log data values, the dataset was processed by min-max normalization and divided into training and testing sets in a 7:3 ratio before the model training. After filtering and removing outliers, SMOTE oversampling processing, and feature selection of the dataset, the grid search method is used to tune the model. After multiple iterations, the optimal classification model is obtained and used in the test well B2 to test the model identification performance. The BSI-XGBoost method was implemented in Python, and the computer hardware used is an Intel CORE 13900K 24-core 36-thread CPU, 96G memory, and NVIDIA RTX4090 24G GPU.

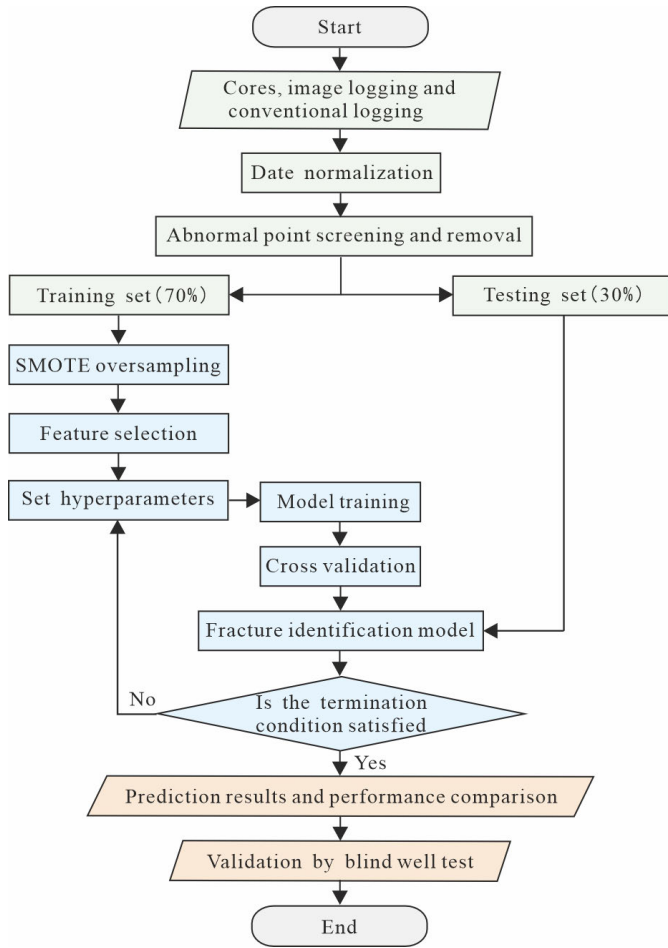


Fig. 4. Schematic flowchart of fracture identification based on the proposed BSI-XGBoost method.

III. EXAMPLES

A. Geological Settings

Mahu Sag is one of the important hydrocarbon-producing sags in the Junggar Basin, with an area of about $5 \times 10^3 \text{ km}^2$. The northwest edge is Zhayier and Halaalate Mountains, while the southwest edge is Zhongguai and Dabasong Uplifts, as shown in Fig. 5(a) and (b). The Mahu Sag was reformed by multiple tectonic deformations during the Late Permian Cenozoic, with a structural style characterized by a southeast dipping monocline with an inclination angle of 3° – 5° [50]. A series of NEE-SWW trending faults is widely distributed in the area, which has a significant impact on the formation of fractures [51]. A set of alkali lake high-quality source rocks are developed in the Fengcheng Formation of Lower Permian, which is mainly composed of multisource mixed fine-grained sediments in deep to semi-deep alkali lake background, with endogenous chemical, volcanic material deposition, and terrigenous clastic depositions [5], [52]. As shown in Fig. 5(c), from bottom to top, the Fengcheng Formation consists of the P_{1f1} , P_{1f2} , and P_{1f3} . Sedimentary rocks are thicker in the northwest and thin out in the southeast, while chemical precipitation first increases and then decreases in the same direction. Sedimentary rocks vary with depth, with sandstone, dolomite, mudstone, tuff, and other rock-type depositions.

B. Characteristics of Natural Fractures

Based on core observations, and image logs interpretation, natural fractures in the study area are highly developed, and could be divided into four groups based on their strikes: north-northwest to south-south-east, northwest-west to southeast-east, near east-west, and near north-south. The average density of fractures is 2.2 m^{-1} , while fractures located on the hanging wall of a thrust fault and close to the fault zone can reach 8 m^{-1} . There are significant differences in the content of brittle minerals in various lithologies of the Fengcheng Formation. Dolomitic rock and tuffaceous fine sandstone have a higher content of brittle minerals, thus a higher density of fractures [53], [54].

Fractures in the Fengcheng Formation are predominantly high-angle shear fractures, which can be categorized into two types as follows.

- 1) Fractures filled with calcareous and siliceous minerals, as shown in Table II(a) and (b). These exhibit white sinusoidal shapes with relatively large amplitude in image logs.
- 2) Open fractures without any mineral precipitation, as shown in Table II(c) and (d). These appear as black sinusoids with relatively large amplitude in image logs.

Additionally, as indicated in Table II(e) and (f), nonfractured sections do not display unique characteristics on image logs; thus, the target layer is labeled based on core inspections and image logs. Fracture identification data are classified into three categories: filled fractures, unfilled (open) fractures, and nonfractured.

C. Logging Responses of Natural Fractures

The differences in the characteristics of the log response between the filled fractures, unfilled fractures, and nonfractured sections, can be distinguished by their amplitude [10]. For example, when drilling in the fractured zone, mud leakage or mud cake thickening is expected, resulting in a smaller value of the CAL. Gamma log (GR) is a radioactive record formed due to the presence of radioactive elements (uranium, thorium, and potassium) that might have been dissolved in the water precipitated in fractures, resulting in higher GR values in filled fractured zones [10], [12]. In open fractures, the P-wave propagates slower, which may lead to an increase in the AC travel time values, compared to the filled fractures [10], [12]. When the fractures are filled with fluid, the density would be generally smaller compared to the filled or nonfractured intervals, resulting in a decrease in the density log (DEN) values [10], [12]. CNL reflects the hydrogen content of the pores. If the fractures are filled with fluid, this will cause an increase in the CNL [11]. In the fractured zone, the formation of water in the open fractures can lead to a decrease in the true formation resistivity log (R_t) values. Therefore, an interval that is characterized by high resistivity values can be a nonfractured zone or a fractured zone that is filled with minerals [12].

The data used for modeling in this study is obtained from a total of eight conventional logging curves from well B1 and the test well B2, including acoustic log (AC), density

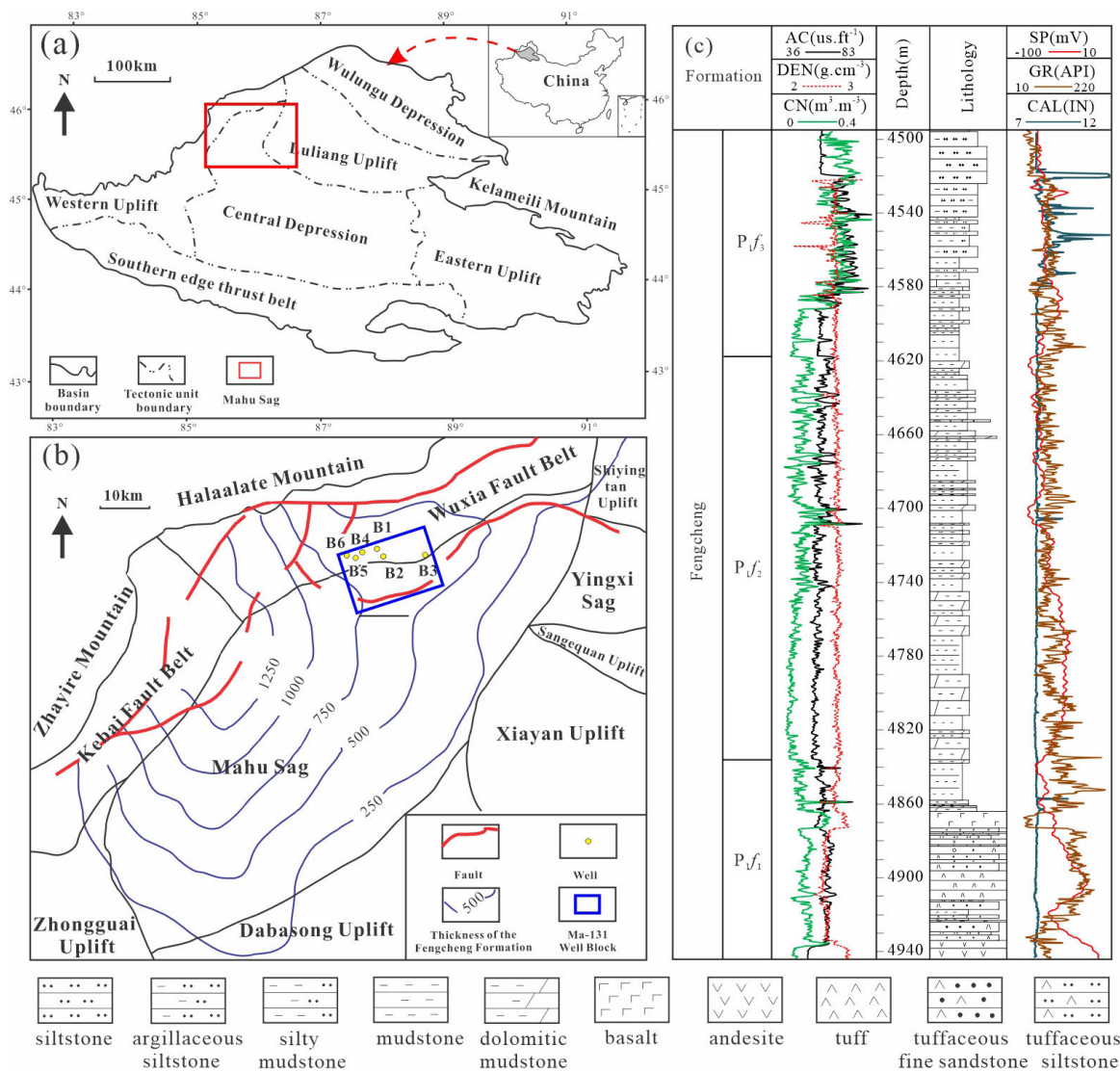


Fig. 5. Tectonic division of the Mahu Sag in the Junggar Basin. (a) Location of the Mahu Sag and Structural Unit of Junggar Basin. (b) Location of the study area and the structural unit of the Mahu Sag. (c) Stratigraphic column of Fengcheng formation.

(DEN), CNL, GR, caliper (CAL), true formation resistivity (R_t), invaded zone resistivity (R_i), and flushed zone resistivity (R_{xo}). Due to the variation in the distribution of natural fractures is mainly influenced by the petrophysical properties of rocks, such as mineral composition, grain size, and geological structural features [55]. These physical properties can cause changes in mechanic properties, which control the development degree of natural fractures [53]. The X-ray diffraction (XRD) analysis results of Fengcheng Formation cores demonstrate that the linear density of high-angle shear fractures increases with larger amounts of brittle mineral (searlesite, feldspar, and dolomite), as shown in Fig. 6. This is because when the content of brittle minerals is higher, the oil shale is more prone to brittle deformation under the same stress conditions [54]. Therefore, in addition to the eight conventional logging curves, the rock brittleness curve obtained from logging interpretation is also used as a part of the model training data.

By inspecting the cored interval (4516–4528 m) of Well B1, three different zones can be identified and compared with the response as having filled fractures, open fractures,

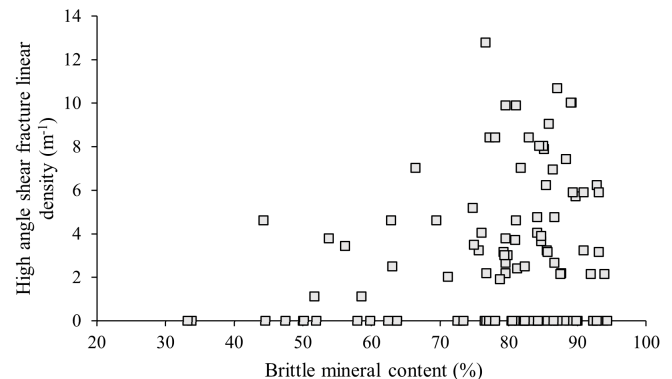

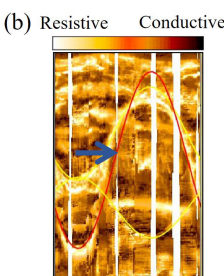
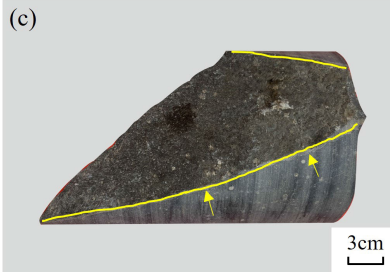
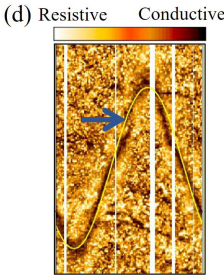

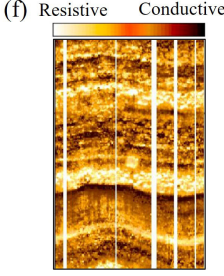


Fig. 6. Schematic showing the relationship between the linear density of high angle shear fractures and brittle mineral content in oil shales of the Fengcheng Formation.

and nonfractured sections. Therefore, unfilled fractures have relatively high AC and CNL values, but lower R_t values while the filled fractures have relatively high GR and low CAL values, as shown in Fig. 7.

TABLE II

PHOTOS OF FRACTURES IN THE CORE SECTION OF FENGCHENG FORMATION IN THE STUDY AREA. (A) FILLED FRACTURE IN CORED SECTION. (B) FILLED FRACTURE IN THE IMAGE LOG. (C) UNFILLED FRACTURE IN CORED SECTION. (D) UNFILLED FRACTURE IN THE IMAGE LOG. (E) NONFRACTURE IN CORED SECTION. (F) NONFRACTURE IN THE IMAGE LOG

Type	Core photos	Image log	Characteristic description
Filled fracture	(a) 	(b) 	(1) Usually composed of calcareous and siliceous minerals. (2) The white sinusoidal shape with relatively large amplitude is observed in image logs.
Unfilled fracture	(c) 	(d) 	(1) Not filled with minerals. (2) The black sinusoidal shape with relatively large amplitude is observed in image logs.
Non-fractured	(e) 	(f) 	No obvious characteristics in core and imaging logs.

Several cross-plots were created based on the conventional logging data and brittleness curves of the Well B1 Fengcheng Formation to better delineate the differences in logging data and curve response in different fractured and nonfracture sections. The overlap and separation of the peaks in the curves reflect the sensitivity of the logging curves to fractures. The AC values of filled fractures are scattered between 54.054 and 78.45 $\mu\text{s}/\text{ft}$, while the brittleness value is between 20.813% and 76.186%. The AC values of unfilled fractures range from 50.742 to 77.766 $\mu\text{s}/\text{ft}$, and the brittleness values range from 18.02% to 82.109%. The distribution of AC values in nonfractured sections ranges from 52.135 to 82.97 $\mu\text{s}/\text{ft}$, and the corresponding brittleness values are found from 16.59% to 87.466%, as shown in Fig. 8(a) and Table III. The CNL values of filled fractures are measured from 0.037 to 0.337 $\text{m}^3\cdot\text{m}^{-3}$, while the DEN values range from 2.421 to 2.746 $\text{g}\cdot\text{cm}^{-3}$. The distribution of CNL values in the unfilled fractured sections is between 0.018 and 0.337 $\text{m}^3\cdot\text{m}^{-3}$, and the corresponding DEN values are found between 2.548 and 2.737 $\text{g}\cdot\text{cm}^{-3}$. The CNL values in the nonfractured sections are between 0.026 and 0.383 $\text{m}^3\cdot\text{m}^{-3}$, while DEN values range from 2.421 to 2.746 $\text{g}\cdot\text{cm}^{-3}$, as shown in Fig. 8(b) and Table III. The GR value of the filled fracture sections is distributed between 43.374 and 165.677 API, and the CAL

value is distributed between 8.352 and 9.023 in The GR value of the unfilled fracture sections is distributed between 58.503 and 160.094 API, and the CAL value is distributed between 8.371 and 9.752 in. The GR value distribution in the nonfractured sections is between 41.269 and 183.043 API, and the CAL value distribution is between 8.332 and 9.665 in, as shown in Fig. 8(c) and Table III. The log (R_t) value of the filled fractured sections is measured between 1.48 and 6.64 log ($\Omega\cdot\text{m}$), and the log (R_i) value between 1.57 and 6.8 log ($\Omega\cdot\text{m}$). The log (R_t) value of the unfilled fractured sections is distributed between 1.7 and 6.7 log ($\Omega\cdot\text{m}$), and the log (R_i) value is found between 1.65 and 6.68 log ($\Omega\cdot\text{m}$). The log (R_t) value in nonfractured segments exhibited a range of 1.15–7.43 log ($\Omega\cdot\text{m}$), while the log (R_i) values vary from 1.17 to 7.58 log ($\Omega\cdot\text{m}$), as shown in Fig. 8(d) and Table III.

The aforementioned results show that the logging values for filled fractures, unfilled, and nonfractured sections in the cross plots of logging curves are vastly different. However, there is still a considerable overlap, which makes it challenging to distinguish and separate these zones, solely based on the logging data. Hence, it is necessary to further establish a nonlinear and intelligent fracture identification model to achieve such a goal with high accuracy.

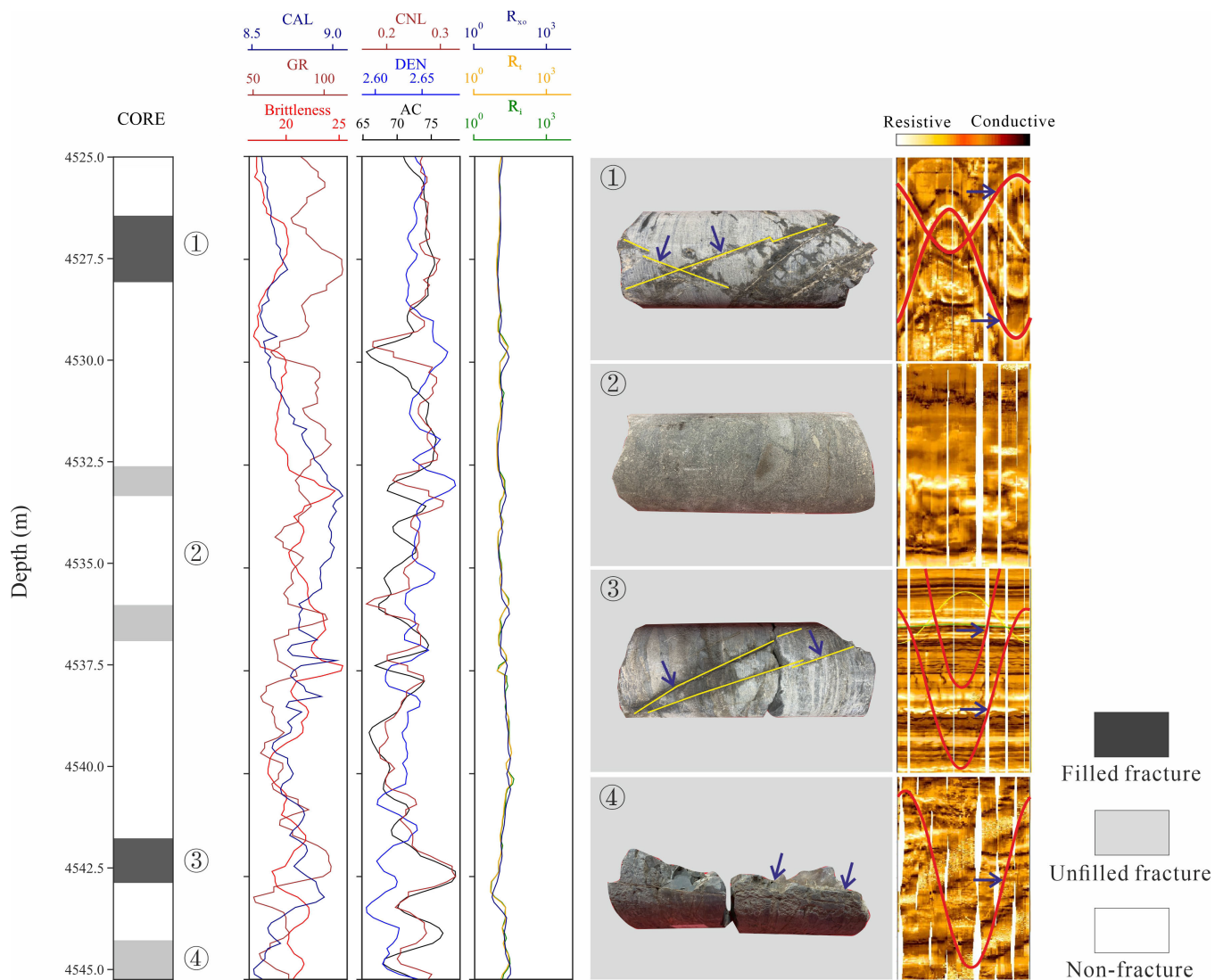


Fig. 7. Fracture distribution and well logging response of Well B1 in the study area. The selected depth segment includes filled fracture, unfilled fracture, and nonfracture segments.

D. Anomaly Point Removal and Feature Selection

Through processing datasets using the iForest, 147 abnormal points were screened from the original data, of which a score less than 0 represents the anomaly, as shown in Fig. 9. On this basis, this study also conducted a feature contribution ranking, as indicated in Fig. 10.

The feature selection process is divided into two steps: first, clarify the ranking of feature contribution. Second, clarify the number of preferred features. The XGBoost method can analyze feature weights, obtain the contribution of each feature on the decision tree, and obtain feature contribution ranking. Since a single decision tree in XGBoost calculates feature importance by improving the performance metrics through each feature splitting point, the closer it is to the root node, the greater the weight will be, thus important features are selected by the tree.

We performed feature selection on nine curves in the training set and the results demonstrate that brittleness has the most contribution to the fracture identification of the model,

followed by CAL, AC, and GR, while the least contribution belongs to $\log(R_t)$ and $\log(R_f)$, as shown in Fig. 10. On this basis, the training data are randomly divided by 7:3 for the second time. Based on this ranking, 1–9 features in their contribution order are selected for training the model. The results indicate that although features such as brittleness, CAL, and AC are the most important to fracture identification, all features participating in the model still provide us with good accuracy in the training data, as indicated in Fig. 11. Moreover, we can observe that removing features with less contribution would reduce the noise but also tend to lose useful information in the model.

E. Hyperparameter Optimization

During the model training process, hyperparameter tuning of machine learning classifiers can improve the model's classification ability. There are empirical and automatic parameter tuning methods for hyperparameter optimization [56]. Because the identification model has multiple hyperparameters, tuning each hyperparameter separately may not necessarily result in

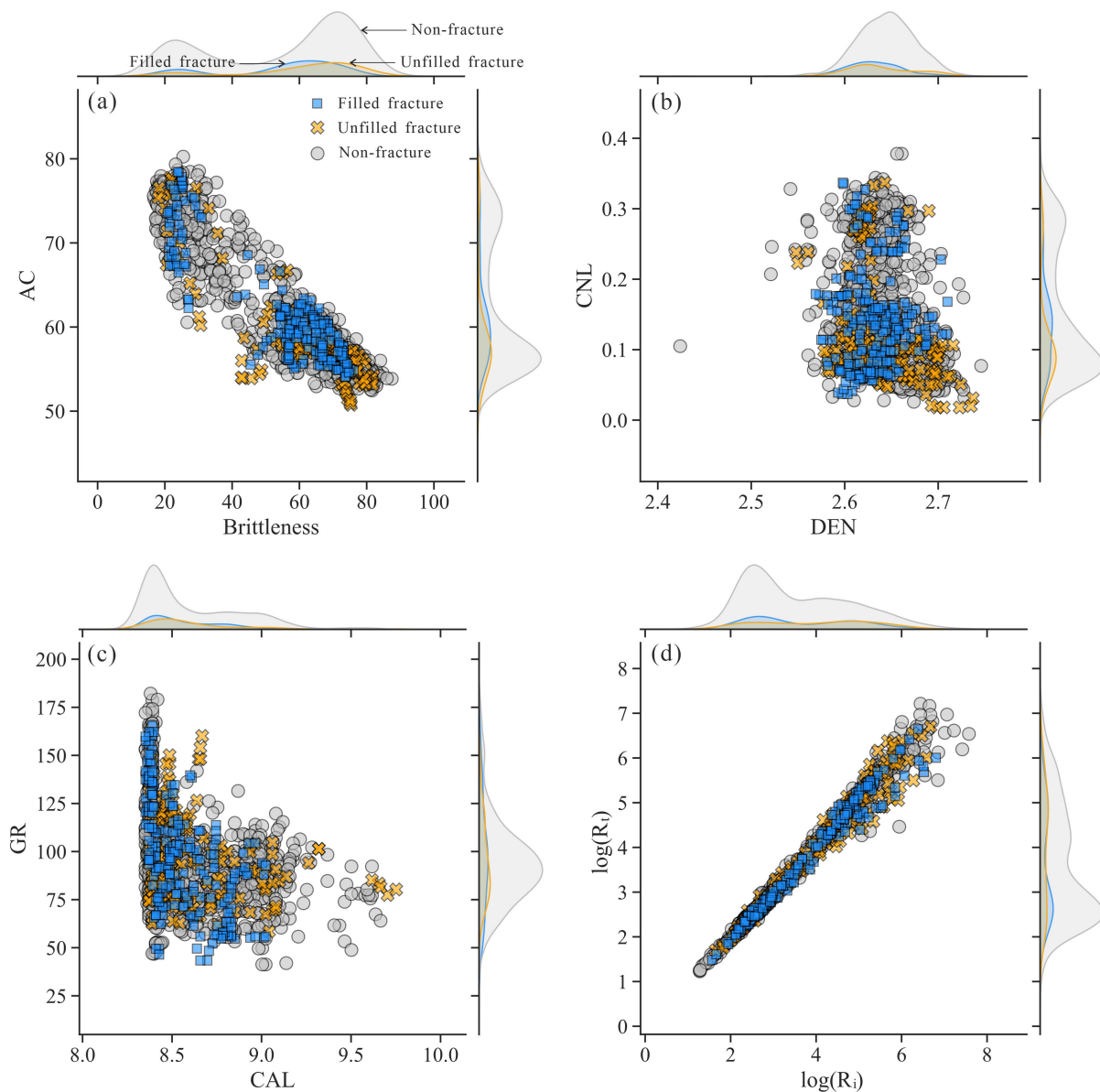


Fig. 8. Cross plots of well logs against fracture labels. (a) AC and Brittleness. (b) CNL and DEN. (c) GR and CAL. (d) $\log(R_t)$ and $\log(R_i)$.

TABLE III
STATISTICS OF CONVENTIONAL WELL-LOGGING DATA IN THE STUDY AREA

	Filled fracture	Unfilled fracture	Non-fractured
<i>Brittleness</i> (%)	20.813 ~ 76.186	18.02 ~ 82.109	16.59 ~ 87.466
<i>CAL</i> (inch)	8.352 ~ 9.023	8.371 ~ 9.752	8.332 ~ 9.665
<i>GR</i> (API)	43.374 ~ 165.677	58.503 ~ 160.094	41.269 ~ 183.043
<i>AC</i> (us.ft ⁻¹)	54.054 ~ 78.45	50.742 ~ 77.766	52.135 ~ 82.97
<i>DEN</i> (g.cm ⁻³)	2.421 ~ 2.746	2.548 ~ 2.737	2.421 ~ 2.746
<i>CNL</i> (m ³ .m ⁻³)	0.037 ~ 0.337	0.018 ~ 0.337	0.026 ~ 0.383
$\log(R_t)$ (log (Ω. m))	1.57 ~ 6.8	1.65 ~ 6.68	1.17 ~ 7.58
$\log(R_i)$ (log (Ω. m))	1.48 ~ 6.64	1.7 ~ 6.7	1.15 ~ 7.43

the best classification performance of the model. Only by tuning multiple hyperparameters simultaneously can we determine the global optimal solution for all hyperparameters [56]. The empirical method not only requires high requirements for

operators but also makes it too difficult to find the global optimal solution solely through human experience. Therefore, this study adopts the grid search tool in the automatic parameter adjustment method to find the optimal value of

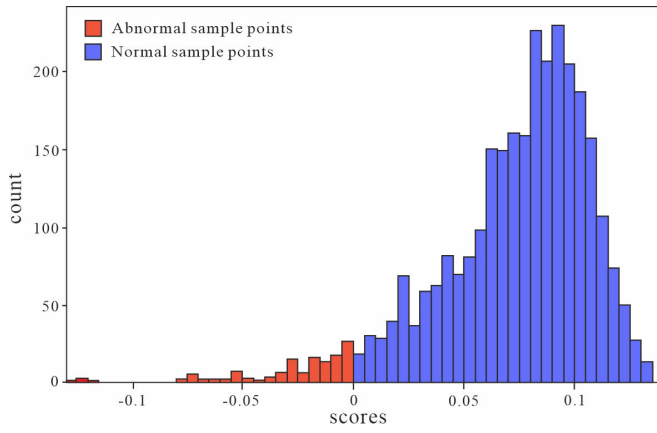


Fig. 9. Histogram of detection results of isolated forest for abnormal data points.

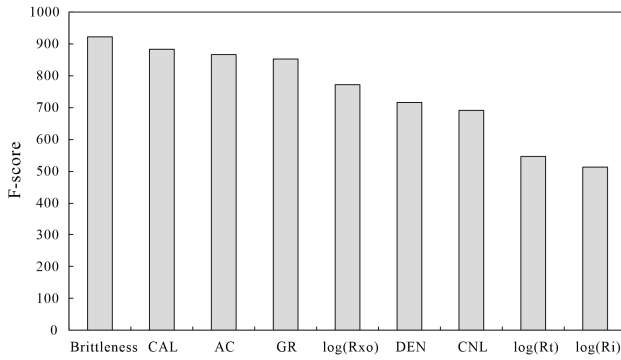


Fig. 10. Ranking of the contribution of nine features to fracture identification model.

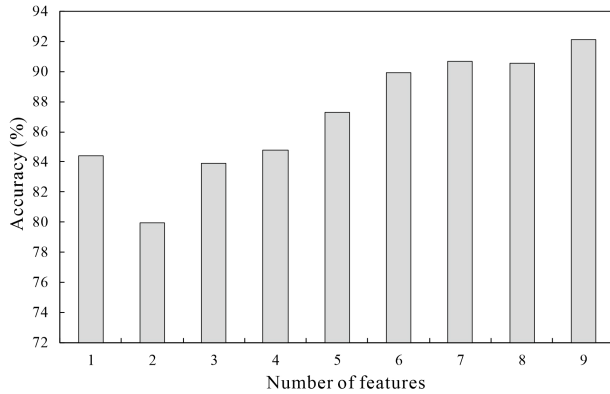


Fig. 11. Relationship between the number of feature selections and classification accuracy.

hyperparameters. The principle is to divide the hyperparameters to be optimized into grids within a certain spatial range, and search for the optimal solution of hyperparameters by traversing all intersections in the grid [10].

The grid search process selects a grid with a smaller step size to optimize the hyperparameters that determine the splitting conditions of the leaf nodes and model complexity in the model, including “gamma,” “maximum depth,” and “min child weight.” Among them, the “gamma” represents the minimum loss reduction required for leaf node splitting,

and the larger this value, the harder it is for leaf nodes to split. The “max depth” represents the maximum depth of the decision tree, and the larger the value, the higher the model complexity. The “min child weight” represents the minimum leaf node weight. If the weight of all samples on the leaf node is less than this value in a split, the split stops. As shown in Fig. 12, an increase in the “gamma” and “min child weight” can lead to a decrease in the classification performance of the model. If the “max depth” is too small, it will lead to underfitting of the model. The optimal hyperparameter set of the model is shown in Table IV, where the global optimal solutions for “gamma,” “max depth,” and “min child weight” are chosen as 0, 7, and 1, respectively.

F. Comparison of Identification Performance With Origin Method

To emphasize the importance of rock brittleness, outlier detection, and data oversampling in BSI-XGBoost, two additional reference models were chosen for comparison. The first model does not incorporate rock brittleness in the dataset and does not apply outlier detection or oversampling techniques, referred to as XGBoost hereafter. The second model includes rock brittleness in the dataset but does not involve outlier detection or oversampling, referred to as B-XGBoost hereafter.

We applied the model to fracture identification in the testing data of Well B1 and the results are shown in Fig. 13(a). In this figure, the red area of the confusion matrix indicates the number of samples incorrectly identified along with the ratio of these samples to the total number of samples in the testing data. The green area represents the number of correctly identified samples with the ratio of such sample numbers to the total number of samples in the testing data. The bottom row of the confusion matrix uses light gray to represent precision (P_r), the rightmost column represents recall (R_e), and the dark gray box at the bottom right corner represents accuracy (A_c). Considering Fig. 13(a), the original XGBoost method identification accuracy is 87.59%. The precision of nonfracture identification is 98.38%, and the recall is 88.7%. The precision of unfilled fracture is 23.81%, and the recall is 68.18%. The precision of filled fractures is 32.88%, and the recall is 72.73%. The B-XGBoost method identification accuracy is 88.72%. The precision of nonfracture identification is 98.52%, and the recall is 89.69%. The precision of unfilled fracture is 31.75%, and the recall is 76.92%. The precision of filled fractures is 38.36%, and the recall is 75.68%, as indicated in Fig. 13(b). Finally, The BSI-XGBoost method identification accuracy is 92.45%. The precision of nonfracture identification is 95.24% and the recall is 96.82%. The precision of unfilled fracture is 70%, and the recall is 51.85%. The precision of filled fracture is 72.88%, and recall is 75.44%, as shown in Fig. 13(c).

To further validate the generalization ability of the models built using these three methods for fracture identification in the test well B2, the 4565–4595 m interval was evaluated. The identification results are presented in Fig. 14, where four distinct depth intervals are depicted to highlight the performance of the identification. The main observations are summarized as follows.

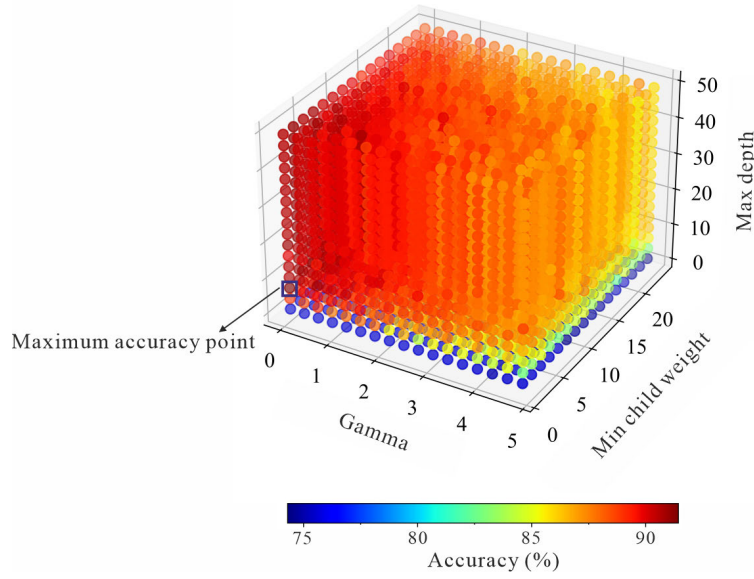


Fig. 12. Grid search process for finding the globally optimal solutions of the three hyperparameters “gamma,” “maximum depth,” and “min child weight” in the fracture identification model.

TABLE IV
HYPERPARAMETER SELECTION AND OPTIMAL SETTING

Hyperparameters	Range	Optimal setting value
Gamma	0-5	0
Max depth	1-50	7
Min child weight	1-25	1

- 1) In the segment ranging from 4565.8 to 4566.9 m, which is a nonfractured section according to the core data, XGBoost identified this interval as a filled fracture. However, B-XGBoost and BSI-XGBoost partially recognized the correct nature of the interval.
- 2) From 4567.8 to 4570.25 m, another nonfractured section was present. BSI-XGBoost comparatively performed well in identifying this section correctly, whereas XGBoost and B-XGBoost misclassified most of this interval as fractures.
- 3) Between 4577.8 and 4580.6 m, a few filling fractures were observed. B-XGBoost failed to recognize this fracture segment entirely, while XGBoost accurately identified the presence of fractures but mislabeled them as unfilled fractures. BSI-XGBoost exhibited the highest accuracy among these three methods in this specific section.
- 4) In the interval from 4591.3 to 4592.2 m, unfilled fractures were developed. XGBoost and B-XGBoost incorrectly identified this section, whereas BSI-XGBoost correctly recognized the presence of fractures.

By comparing the results, it is evident that BSI-XGBoost achieves the highest identification accuracy (92.45%) among the three models in the testing data. Furthermore, BSI-XGBoost demonstrates slightly superior precision and recall for both fractured and nonfractured segments compared to the other methods. Additionally, BSI-XGBoost exhibits better consistency with the core data in the section of the test

well, in contrast to the original XGBoost and B-XGBoost models.

IV. DISCUSSION

A. Influence of SMOTE on Fracture Identification Results

Based on core inspections, the length of the fractured section in the wellbore is always much less than that of the nonfractured section. Therefore, we expect to encounter more problems with labeling nonfractured sections compared to fractured sections in the data used for model training. The ratio of nonfractured to unfilled and filled fractures in Wells B1 and B2 is approximately 11:1:1. If there is a problem of data imbalance in the model training process, model training tends to focus more on nonfractured with a large number of samples, while neglecting filled and open fractures with a small number, resulting in generally low identification precision and recall for filled and open fractures. Therefore, in the data preprocessing phase, we use the SMOTE algorithm [42] to oversample the fracture data, so that the number of fracture samples is about 0.8 times the number of nonfractures, to overcome the problem of low precision in identifying fractured sections caused by the imbalance data. As shown in Table V, the accuracy of the model testing data trained based on the original data is 90.77%. The precision of the nonfractures, unfilled fractured, and filled fractures section is 99.04%, 25%, and 43.64%, respectively while the recall is 91.84%, 68.42%, and 77.42%, respectively. The accuracy of the model trained by the SMOTE algorithm after oversampling is 92.45%. The precision of the nonfractured, unfilled fractured, and filled fractured sections is 95.24%, 70%, and 72.88%, respectively, with the recall values that are 96.82%, 51.85% and 75.44%, respectively. Through comparison, although the accuracy of the model trained by the original data is 90.77%, the accuracy of both filled and unfilled fractures is less than 45%, indicating that the model cannot accurately identify fractured intervals. Compared with the original data, the accuracy of the oversampling data training

True label	0	(a) 730 83.14%	48 5.47%	45 5.13%	88.7% 11.3%
	1	3 0.34%	15 1.71%	4 0.46%	68.18% 31.82%
	2	9 1.03%	0 0.0%	24 2.73%	72.73% 27.27%
		98.38% 1.62%	23.81% 76.19%	32.88% 67.12%	87.59% 12.41%
		0	1	2	Predicted label

True label	0	(b) 731 83.26%	43 4.9%	41 4.67%	89.69% 10.31%
	1	2 0.23%	20 2.28%	4 0.46%	76.92% 23.08%
	2	9 1.03%	0 0.0%	28 3.19%	75.68% 24.32%
		98.52% 1.48%	31.75% 68.25%	38.36% 61.64%	88.72% 11.28%
		0	1	2	Predicted label

True label	0	(c) 700 83.93%	10 1.2%	13 1.56%	96.82% 3.18%
	1	23 2.76%	28 3.36%	3 0.36%	51.85% 48.15%
	2	12 1.44%	2 0.24%	43 5.16%	75.44% 24.56%
		95.24% 4.76%	70.0% 30.0%	72.88% 27.12%	92.45% 7.55%
		0	1	2	Predicted label

Fig. 13. Confusion matrix of the testing data of the Well B1 for fracture identification. (a) XGBoost. (b) B-XGBoost. (c) BSI-XGBoost. Nonfracture, unfilled fracture, and filled fracture are represented by 0, 1, and 2.

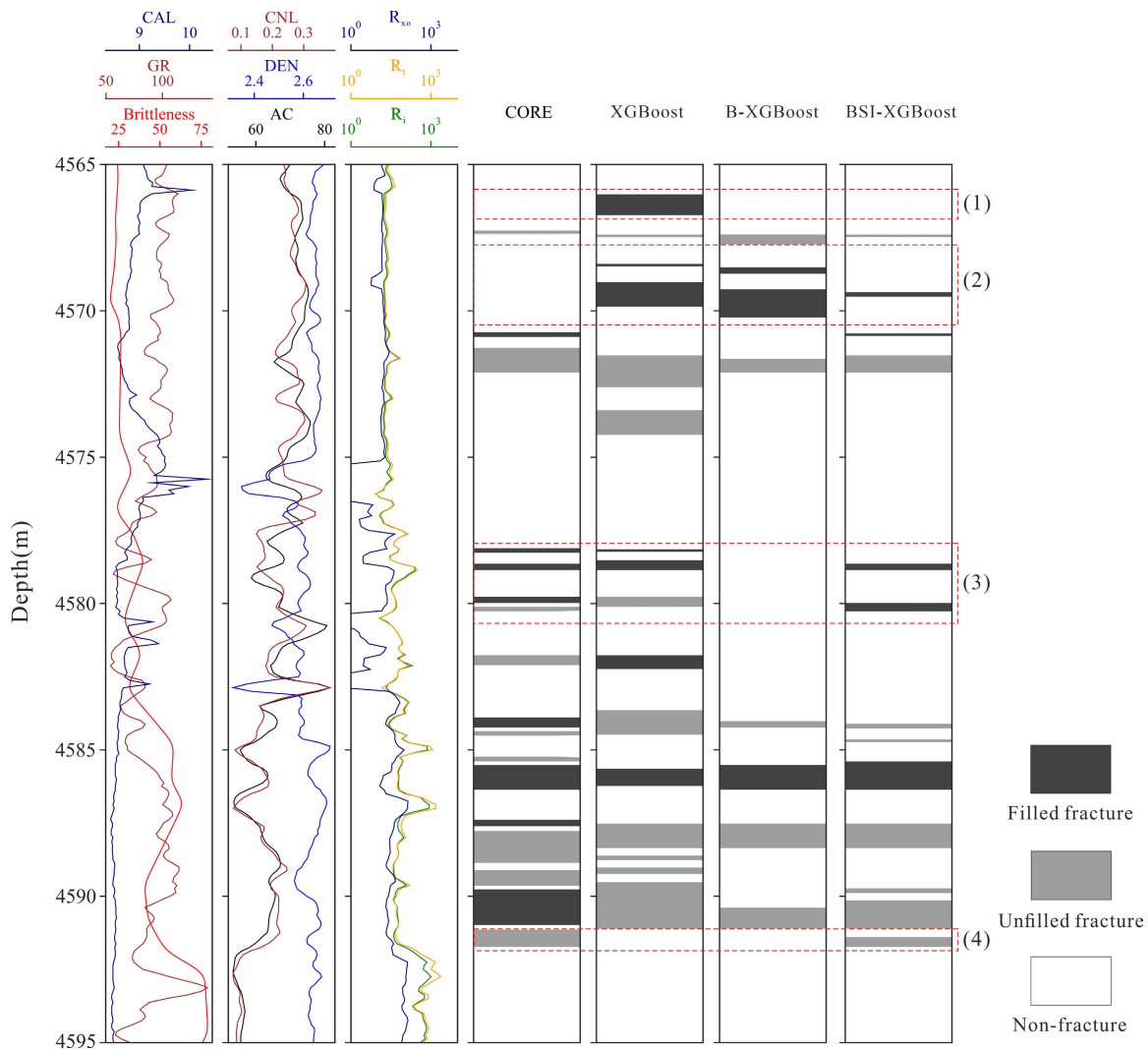


Fig. 14. Comparison of intelligent fracture identification results of the test well using XGBoost, B-XGBoost, and BSI-XGBoost methods with cores.

model is 1.68% higher. The precision of filled and unfilled fractures has significantly improved, signifying the necessity of using the SMOTE algorithm to process training data.

B. Uncertainty Analysis

In the task of fracture identification, uncertainty analysis using a 95% confidence interval can evaluate the reliability

of the predictive results. To analyze the uncertainty in the predictions made by BSI-XGBoost and comparison methods, three approaches were utilized to identify fractures in the Well B1 test set and calculate the 95% confidence intervals for each evaluation metric. By conducting 1000 bootstrap resamplings of the original test dataset, where each resampling contained the same number of instances as the original dataset, multiple new test datasets were generated for the calculation

TABLE V
COMPARISON OF IDENTIFICATION RESULTS BETWEEN
ORIGINAL DATA AND OVERSAMPLED DATA

	Accuracy	Precision	Recall	Type
No smote	90.77%	99.04%	91.84%	Non-fracture
		25%	68.42%	Unfilled fracture
		43.64%	77.42%	Filled fracture
Smote	92.45%	95.24%	96.82%	Non-fracture
		70%	51.85%	Unfilled fracture
		72.88%	75.44%	Filled fracture

of the 95% confidence intervals. The results are presented in Table VI.

The original XGBoost method's identification accuracy was distributed within the 95% confidence interval between 85.54% and 89.53%. The precision of nonfracture identification ranged from 97.44% to 99.19%, and the recall from 86.51% to 90.65%. The precision for unfilled fractures was between 12.3% and 35.63%, with the recall ranging from 52.2% to 82.62%. For filled fractures, the precision was between 21.91% and 44.47%, and the recall ranged from 58.11% to 85.6%. The B-XGBoost method showed identification accuracy within the 95% confidence interval from 86.45% to 90.89%. The precision for nonfracture identification ranged from 97.6% to 99.21%, with the recall from 87.53% to 91.86%. The precision for unfilled fractures was between 20.94% and 44.01%, with the recall from 58.38% to 92.73%. For filled fractures, the precision ranged from 27.9% to 49.32%, and the recall from 59.88% to 89.92%. Finally, the BSI-XGBoost method's identification accuracy fell within the 95% confidence interval between 91.15% and 93.5%. The precision for nonfracture identification ranged from 94.56% to 95.78%, with the recall from 95.1% to 98.37%. The precision for unfilled fractures was between 59.88% and 80.85%, and the recall was from 37.93% to 64.52%. For filled fractures, the precision ranged from 63.97% to 82.03%, with the recall between 65.44% and 82.11%.

In this study, the width of the 95% confidence interval was determined by the difference between the upper and lower limits. A narrower width indicates less uncertainty in the results, suggesting more precise model predictions. Given that each type of fracture identification has corresponding recall and precision, and there are three types in total, the identification results include three measures of precision, three of recall, and one of accuracy. The widths of the 95% confidence intervals for the three methods are depicted in Fig. 15. Notably, the precision and recall for nonfracture samples, as well as the overall accuracy, have smaller 95% confidence interval widths. This is primarily due to the larger number of nonfracture samples in the unbalanced data, which results in relatively lower uncertainty in the model predictions for nonfracture samples. Conversely, for filled and unfilled fractures, the 95% confidence intervals for the three methods are significantly wider, indicating that the fewer number of fracture samples in the unbalanced data leads to relatively greater uncertainty in the model predictions for fracture samples. A horizontal comparative analysis among the three methods shows that the BSI-XGBoost method, compared to XGBoost

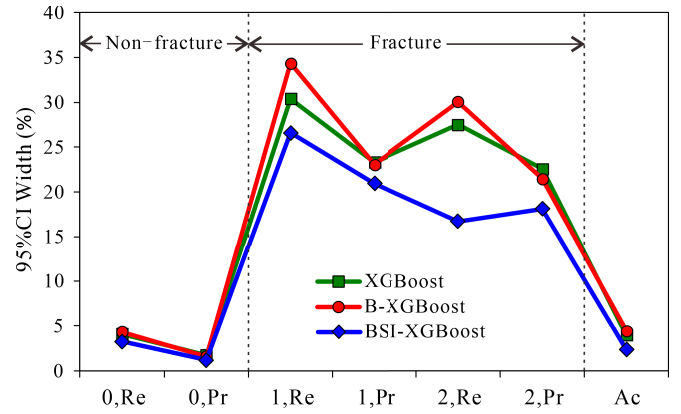


Fig. 15. Comparison of the width of the 95% confidence interval across different types. Nonfracture, unfilled fracture, and filled fracture are represented by 0, 1, and 2. Recall, precision, and accuracy are represented by Re, Pr, and Ac.

and B-XGBoost, generally exhibits the least uncertainty in prediction results, especially for filled and unfilled fractures. The 95% confidence interval widths for recall and precision are noticeably smaller with BSI-XGBoost than with the other two methods, indicating that BSI-XGBoost has relatively lower uncertainty in predicting fracture samples and better identification capabilities for types with fewer samples in unbalanced data.

C. Strengths and Limitations

BSI-XGBoost, an enhancement of the original XGBoost method, tackles the challenges associated with imbalanced datasets and anomalous data points in the training of fracture identification models. This method significantly improves the accuracy of identifying natural fractures. Additionally, it demonstrates superior identification capabilities for fracture samples with fewer instances in unbalanced data, and the widths of the 95% confidence intervals for precision and recall for both filled and unfilled fractures are relatively narrower, indicating a smaller degree of uncertainty in the identification results.

Furthermore, conventional logging for fracture identification presents a highly complex challenge. Applying this model in other areas primarily faces two challenges as follows.

- 1) Due to regional variations in sedimentary environments, lithofacies characteristics, pore fluids, and drilling mud pressures, the conventional logging response characteristics of natural fractures also differ.
- 2) There is no unified standard for logging curves across different regions, making it difficult to directly apply this model elsewhere.

This difficulty is not due to a weak generalization capability of the identification model established by this method, but rather due to significant differences in the original training data from different study areas. Consequently, this method is applicable primarily to the identification of natural fractures in reservoirs that share similar characteristics with those of the Fengcheng Formation in the Mahu Sag. To extend this methodology to other study areas with substantial geological

TABLE VI
IDENTIFICATION RESULTS OF THE WELL B1 TESTING DATA. THE 95% CONFIDENCE INTERVALS
ARE PRESENTED IN THE BRACKETS, RESPECTIVELY

Method	Accuracy	Precision	Recall	Type
XGBoost	87.59% (85.54% - 89.53%)	98.38% (97.44% - 99.19%)	88.7% (86.51% - 90.65%)	Non-fracture
		23.81% (12.3% - 35.63%)	68.18% (52.2% - 82.62%)	Unfilled fracture
		32.88% (21.91% - 44.47%)	72.73% (58.11% - 85.6%)	Filled fracture
B-XGBoost	88.72% (86.45% - 90.89%)	98.52% (97.6% - 99.21%)	89.69% (87.53% - 91.86%)	Non-fracture
		31.75% (20.94% - 44.01%)	76.92% (58.38% - 92.73%)	Unfilled fracture
		38.36% (27.9% - 49.32%)	75.68% (59.88% - 89.92%)	Filled fracture
BSI-XGBoost	92.45% (91.15% - 93.5%)	95.24% (94.56% - 95.78%)	96.82% (95.1% - 98.37%)	Non-fracture
		70.00% (59.88% - 80.85%)	51.85% (37.93% - 64.52%)	Unfilled fracture
		72.88% (63.97% - 82.03%)	75.44% (65.44% - 82.11%)	Filled fracture

differences from the Fengcheng Formation, it is crucial to incorporate additional training data from those areas, including other interpretive curves relevant to fracture development in the analysis, and establish new fracture identification models based on the training dataset of new study areas.

V. CONCLUSION

To overcome the challenges in fracture identification in oil shale based on easily accessible conventional logging data, we have proposed an integrated ensemble learning method (BSI-XGBoost). BSI-XGBoost uses core observation and image logging as labels to delineate the nonlinear relationship between fracture presence and conventional logging response, to accurately identify the fractured intervals. In addition, BSI-XGBoost integrates geological knowledge into the model construction process by incorporating rock brittleness, which is one of the crucial factors influencing fracture development. This methodology also successfully tackles the challenge of imbalanced data, where the number of nonfracture samples far exceeds the number of fracture samples in fracture identification. More importantly, BSI-XGBoost seamlessly integrates the iForest method and feature selection into the model construction process, successfully mitigating the problem of amplified error interference due to noise during the iterative boosting process. Therefore, BSI-XGBoost greatly enhances the robustness and generalizability of the fracture identification model. Results of fracture identification using BSI-XGBoost demonstrate an impressive accuracy of 92.45% on the testing set. This achievement showcases a significant improvement of 4.86% and 3.73% over the original XGBoost and B-XGBoost models, respectively. These findings underscore the superiority of BSI-XGBoost in intelligently identifying fractures by leveraging conventional logging techniques.

REFERENCES

- [1] X. Liu et al., "Deep classified autoencoder for lithofacies identification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, doi: [10.1109/TGRS.2021.3139931](https://doi.org/10.1109/TGRS.2021.3139931).
- [2] Y. Chen and D. Zhang, "Physics-constrained deep learning of geomechanical logs," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5932–5943, Aug. 2020.
- [3] L. Gong et al., "Characterization, controlling factors and evolution of fracture effectiveness in shale oil reservoirs," *J. Petroleum Sci. Eng.*, vol. 203, Aug. 2021, Art. no. 108655.
- [4] G. Huang, X. Chen, C. Luo, and Y. Chen, "Geological structure-guided initial model building for prestack AVO/AVA inversion," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1784–1793, Feb. 2021.
- [5] G. Lu et al., "Lithology identification using graph neural network in continental shale oil reservoirs: A case study in mahu sag, Junggar basin, western China," *Mar. Petroleum Geol.*, vol. 150, Apr. 2023, Art. no. 106168.
- [6] Y. Tang, J. Cao, W.-J. He, X.-G. Guo, K.-B. Zhao, and W.-W. Li, "Discovery of shale oil in alkaline lacustrine basins: The late Paleozoic Fengcheng Formation, Mahu Sag, Junggar Basin, China," *Petroleum Sci.*, vol. 18, no. 5, pp. 1281–1293, Oct. 2021.
- [7] G. Liu et al., "Natural fractures in tight gas volcanic reservoirs and their influences on production in the Xujiaweizi depression, Songliao basin, China," *AAPG Bull.*, vol. 104, no. 10, pp. 2099–2123, Oct. 2020.
- [8] L. Zeng, J. Jiang, and Y. Yang, "Fractures in the low porosity and ultra-low permeability glutenite reservoirs: A case study of the late eocene Hetaoyuan formation in the Anpeng oilfield, Nanxiang basin, China," *Mar. Petroleum Geol.*, vol. 27, no. 7, pp. 1642–1650, Aug. 2010.
- [9] L. Yang et al., "High-fidelity permeability and porosity prediction using deep learning with the self-attention mechanism," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 7, pp. 3429–3443, Jul. 2023.
- [10] S. Dong et al., "Fracture identification by semi-supervised learning using conventional logs in tight sandstones of Ordos basin, China," *J. Natural Gas Sci. Eng.*, vol. 76, Apr. 2020, Art. no. 103131.
- [11] A. Ja'fari, A. Kadkhodaie-Ilkhchi, Y. Sharghi, and K. Ghanavati, "Fracture density estimation from petrophysical log data using the adaptive neuro-fuzzy inference system," *J. Geophys. Eng.*, vol. 9, no. 1, pp. 105–114, Feb. 2012.
- [12] B. Tokhmechi, H. Memarian, H. A. Noubari, and B. Moshiri, "A novel approach proposed for fractured zone detection using petrophysical logs," *J. Geophys. Eng.*, vol. 6, no. 4, pp. 365–373, Dec. 2009.
- [13] Y. Wu et al., "Robust unilateral alignment for subsurface lithofacies classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, doi: [10.1109/TGRS.2021.3070050](https://doi.org/10.1109/TGRS.2021.3070050).
- [14] K. Gao, L. Huang, and Y. Zheng, "Fault detection on seismic structural images using a nested residual U-Net," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2021.
- [15] H. Liu et al., "Rock thin-section analysis and identification based on artificial intelligent technique," *Petroleum Sci.*, vol. 19, no. 4, pp. 1605–1621, Aug. 2022.
- [16] X.-Y. Liu, L. Zhou, X.-H. Chen, and J.-Y. Li, "Lithofacies identification using support vector machine based on local deep multi-kernel learning," *Petroleum Sci.*, vol. 17, no. 4, pp. 954–966, Aug. 2020.
- [17] S. Li et al., "Deep-learning inversion of seismic data," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2135–2149, Mar. 2020.
- [18] N. Liu, T. Huang, J. Gao, Z. Xu, D. Wang, and F. Li, "Quantum-enhanced deep learning-based lithology interpretation from well logs," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2021.
- [19] O. M. Saad, Y. Chen, A. Savvaidis, W. Chen, F. Zhang, and Y. Chen, "Unsupervised deep learning for single-channel earthquake data denoising and its applications in event detection and fully automatic location," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, doi: [10.1109/TGRS.2022.3209932](https://doi.org/10.1109/TGRS.2022.3209932).
- [20] O. M. Saad et al., "EQCCT: A production-ready earthquake detection and phase-picking method using the compact convolutional transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4507015.
- [21] S. Tewari and U. D. Dwivedi, "Ensemble-based big data analytics of lithofacies for automatic development of petroleum reservoirs," *Comput. Ind. Eng.*, vol. 128, pp. 937–947, Feb. 2019.
- [22] Y. Chen, O. M. Saad, A. Savvaidis, Y. Chen, and S. Fomel, "3D micro-seismic monitoring using machine learning," *J. Geophys. Res., Solid Earth*, vol. 127, no. 3, p. 2021, Mar. 2022, Art. no. e2021JB023842.

- [23] Y. Chen, A. Savvaiddis, S. Fomel, O. M. Saad, and Y. Chen, "RFloc3D: A machine-learning method for 3-D microseismic source location using P- and S-wave arrivals," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, doi: [10.1109/TGRS.2023.3236572](https://doi.org/10.1109/TGRS.2023.3236572).
- [24] G. F. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Mach. Learn.*, vol. 9, no. 4, pp. 309–347, Oct. 1992.
- [25] S. Bhattacharya and S. Mishra, "Applications of machine learning for facies and fracture prediction using Bayesian network theory and random forest: Case studies from the Appalachian basin, USA," *J. Petroleum Sci. Eng.*, vol. 170, pp. 1005–1017, Nov. 2018.
- [26] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, "Learning K for kNN classification," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 3, pp. 1–19, 2017.
- [27] H. He, "Fracture identification in conventional log through KNN classification algorithm based on slope of logging curve: A case study of reef flat facies reservoir in puguang gas field," *Sino-Global Energy*, vol. 19, no. 1, pp. 70–74, 2014.
- [28] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
- [29] T. Li, R. Wang, Z. Wang, M. Zhao, and L. Li, "Prediction of fracture density using genetic algorithm support vector machine based on acoustic logging data," *Geophysics*, vol. 83, no. 2, pp. D49–D60, Mar. 2018.
- [30] G.-R. Shi, "Superiorities of support vector machine in fracture prediction and gassiness evaluation," *Petroleum Explor. Develop.*, vol. 35, no. 5, pp. 588–594, Oct. 2008.
- [31] S. Fei et al., "Assessment of ensemble learning to predict wheat grain yield based on UAV-multispectral reflectance," *Remote Sens.*, vol. 13, no. 12, p. 2338, Jun. 2021.
- [32] Y. Lu, Z. Zhang, D. Shanguan, and J. Yang, "Novel machine learning method integrating ensemble learning and deep learning for mapping debris-covered glaciers," *Remote Sens.*, vol. 13, no. 13, p. 2595, Jul. 2021.
- [33] G. Lu et al., "Fracture identification based on graph pooling and graph construction in continental shale," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, doi: [10.1109/TGRS.2024.3397861](https://doi.org/10.1109/TGRS.2024.3397861).
- [34] G. Lu et al., "Bedding-parallel fracture density prediction using graph convolutional network in continental shale oil reservoirs: A case study in Mahu Sag, Junggar basin, China," *Mar. Petroleum Geol.*, vol. 167, Sep. 2024, Art. no. 106992.
- [35] S.-Q. Dong et al., "How to improve machine learning models for lithofacies identification by practical and novel ensemble strategy and principles," *Petroleum Sci.*, vol. 20, no. 2, pp. 733–752, Apr. 2023.
- [36] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, and Y. Zhou, "A novel ensemble method for classifying imbalanced data," *Pattern Recognit.*, vol. 48, no. 5, pp. 1623–1637, 2015.
- [37] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.
- [38] R. E. Schapire, "The strength of weak learnability," *Mach. Learn.*, vol. 5, pp. 197–227, Jun. 1990.
- [39] M. Opitz, G. Waltner, H. Possegger, and H. Bischof, "Deep metric learning with BIER: Boosting independent embeddings robustly," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 276–290, Feb. 2020.
- [40] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers Neurobot.*, vol. 7, p. 21, Jan. 2013.
- [41] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining*, Aug. 2008, pp. 413–422.
- [42] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
- [43] O. M. Saad et al., "SCALODEEP: A highly generalized deep learning framework for real-time earthquake detection," *J. Geophys. Res., Solid Earth*, vol. 126, no. 4, p. 2020, Apr. 2021.
- [44] O. M. Saad, Y. Chen, A. Savvaiddis, S. Fomel, and Y. Chen, "Real-time earthquake detection and magnitude estimation using vision transformer," *J. Geophys. Res., Solid Earth*, vol. 127, no. 5, May 2022, Art. no. e2021JB023657.
- [45] S. Ma et al., "Fault damage zone and its effect on deep shale gas: Insights from 3D seismic interpretation in the southern Sichuan basin, China," *J. Struct. Geol.*, vol. 170, May 2023, Art. no. 104848.
- [46] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
- [47] T. Chen, "XGBoost: eXtreme gradient boosting," *R Package Version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.
- [48] T. S. Bressan, M. Kehl de Souza, T. J. Girelli, and F. C. Junior, "Evaluation of machine learning methods for lithology classification using geophysical data," *Comput. Geosci.*, vol. 139, Jun. 2020, Art. no. 104475.
- [49] J. Shi, X. Zhao, L. Zeng, Y. Zhang, and S. Dong, "Identification of coal structures by semi-supervised learning based on limited labeled logging data," *Fuel*, vol. 337, Apr. 2023, Art. no. 127191.
- [50] W. Dang et al., "Genesis and distribution of oils in Mahu Sag, Junggar basin, NW China," *Petroleum Exploration Develop.*, vol. 50, no. 4, pp. 840–850, Aug. 2023.
- [51] G. Liu et al., "Distribution pattern of natural fractures in lacustrine shales: A case study of the Fengcheng formation in the Mahu Sag of the Junggar basin, China," *Frontiers Earth Sci.*, vol. 11, May 2023, Art. no. 1207033.
- [52] X. Wang et al., "Multi-scale natural fracture prediction in continental shale oil reservoirs: A case study of the Fengcheng formation in the Mahu Sag, Junggar basin, China," *Frontiers Earth Sci.*, vol. 10, May 2022, Art. no. 929467.
- [53] G. Liu et al., "Natural fractures in deep continental shale oil reservoirs: A case study from the Permian Lucaogou formation in the eastern Junggar basin, Northwest China," *J. Struct. Geol.*, vol. 174, Sep. 2023, Art. no. 104913.
- [54] L. Zeng et al., "Natural fractures and their influence on shale gas enrichment in Sichuan basin, China," *J. Natural Gas Sci. Eng.*, vol. 30, pp. 1–9, Mar. 2016.
- [55] J. C. Lorenz, J. L. Sterling, and D. S. Schechter, "Natural fractures in the Spraberry Formation, Midland basin, Texas: The effects of mechanical stratigraphy on fracture variability and reservoir behavior," *AAPG Bull.*, vol. 86, pp. 505–524, Jan. 2002.
- [56] Y. Xie, C. Zhu, W. Zhou, Z. Li, X. Liu, and M. Tu, "Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances," *J. Petroleum Sci. Eng.*, vol. 160, pp. 182–193, Jan. 2018.