Earth Science Informatics
https://doi.org/10.1007/s12145-024-01450-z

RESEARCH

Application of SVC, k-NN, and LDA machine learning algorithms for improved prediction of Bioturbation: Example from the Subei Basin, China

Jonathan Atuquaye Quaye[1] · Kwame Sarkodie[1] · Zaixing Jiang[2] · Chenlin Hu[3] · Joshua Agbanu[1] · Stephen Adjei[1] · Baiqiang Li[4]

Received: 8 June 2024 / Accepted: 9 August 2024

Abstract

Three supervised machine learning (ML) classification algorithms: Support Vector Classifier (SVC), K- Nearest Neighbour (K-NN), and Linear Discriminant Analysis (LDA) classification algorithms are combined with seventy-six (76) data points of nine (9) core sample datasets retrieved from five (5) selected wells in oilfields of the Subei Basin to delineate bioturbation. Application of feature selection via p-score and f-scoring reduced the number of relevant features to 7 out of the 12 considered. Each classifier underwent model training and testing allocating 80% of the data for training and the remaining 20% for testing. Under the model training, optimization of hyperparameters of the SVC (C, Gamma and Kernel) and K-NN (K value) was performed via the grid search to understand the best form of the decision boundaries that provides optimal accuracy of prediction of Bioturbation. Results aided the selection of optimized SVC hyperparameters such as a linear kernel, C-1000 and Gamma parameter—0.10 that provided a training accuracy of 96.17%. The optimized KNN classifier was obtained based on the K = 5 nearest neighbour to obtain a training accuracy of 73.28%. The training accuracy of the LDA classifier was 67.36% which made it the worst-performing classifier in this work. Further cross-validation based on a fivefold stratification was performed on each classifier to ascertain model generalization and stability for the prediction of unseen test data. Results of the test performance of each classifier indicated that the SVC was the best predictor of the bioturbation index at 92.86% accuracy, followed by the K-NN model at 90.48%, and then the LDA classifier which gave the lowest test accuracy at 76.2%. The results of this work indicate that bioturbation can be predicted via ML methods which is a more efficient and effective means of rock characterization compared to conventional methods used in the oil and gas industry.

Keywords Machine learning · Bioturbation · Supervised learning · SVC · K-NN, LDA

Introduction

Bioturbation is a geological phenomenon where the displacement of sediments is attained via the activity of living organisms. This phenomenon can alter not only the textural but also the petrophysical characteristics of a rock (Bromley 1996). Past research conducted on diverse formations within the Subei region has confirmed the significant impact of bioturbation on reservoir quality, underscoring its crucial role in determining a reservoir's productivity. These studies identified ignored secondary reservoir targets that may have increased reserve estimates in petroleum fields (Quaye et al. 2019, 2022, 2023). Analyzing textural parameters within sedimentary rocks poses significant challenges, particularly complex bioturbation. This complexity arises from the wide variation in borings, rootlets, size and intricacy of

Communicated by: Hassan Babaie

✉ Jonathan Atuquaye Quaye
jonathanatuquayequaye@gmail.com

[1] Department of Petroleum Engineering, Faculty of Civil and Geo-Engineering, College of Engineering, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

[2] School of Energy Resources, China University of Geosciences Beijing, No. 29 Xueyuan Road, Haidian, Beijing 100083, China

[3] School of Geology and Mining Engineering, Xinjiang University, Urumqi 830017, China

[4] School of Earth and Environment, Anhui University of Science and Technology, Huainan 232001, China

Published online: 28 August 2024

burrows, and the regular and quick vertical and horizontal alterations, often due to factors like crosscutting burrows and the arrangement of trace-forming endobenthic communities (Bromley and Ekdale 1984). Over time, researchers have adopted various approaches to tackle the issue of describing bioturbation. These methods range from early classification attempts (Schäfer 1956), semi-quantitative measurements proposed by Reineck (1963), quantitative estimations outlined by Dorador et al. (2014), to mathematical modelling as discussed by Guinasso and Schink (1975).

Bioturbation has over the long term greatly affected reservoir quality through modifications on porosity, permeability, or their effects upon depositional stability (Pemberton & Gingras 2005) Burrowing can increase porosity and permeability by opening pathways among grains or reduce these properties if the resulting burrow compacts bordering sediments with finer infill material (Tonkin et al. 2010). Bioturbation impairs depositional interpretations, homogenizes the microstratigraphic distribution of sediment layers, and affects redox chemistry by mixing oxygenated surface sediments with subsurface reducing zones. It also influences compaction and cementation which can either preserve or modify porosity and permeability. In general, bioturbation causes heterogeneity at different scales that can affect fluid flow and reservoir performance making it a key factor to consider in an efficient hydrocarbon recovery (Gingras et al. 1999; Hovikoski et al. 2008).

Machine learning, a component of artificial intelligence, comprises diverse data processing approaches like classification, regression, and clustering. It can be categorized into supervised and unsupervised techniques, delineating two primary branches within this field (Hall 2016). Supervised learning in artificial intelligence involves training a computer algorithm using labelled input data (herein training wells data) to predict specific outputs. Through iterative training, the algorithm learns to recognize hidden patterns and connections between the input and output data, ultimately allowing it to provide precise predictions when given new, unlabelled data (herein test dataset) (Mohri et al. 2012). Mandal and Rezaee (2019) asserted that the use of machine learning, especially when combined with wells data, has gained significant traction in tackling geoscientific issues within sectors of the oil and gas industry. Its application has been widely used in the analysis of various geological, geophysical and petrophysical characterizations (Deshenenkov and Polo 2020; Gharavi et al. 2022; Hansen et al. 2023; Mohammadinia et al. 2023). Fomel & Liu (2017) interpreted seismic data to identify subsurface geological structures and predict reservoir properties. Machine learning models can classify minerals and rocks based on their spectral signatures obtained from remote sensing data (Crosta and Souza Filho 1998). Sarma and Gupta (2000) ("Application of Neural Networks to Tunnel Data Analysis," 1998) used machine learning for reservoir characterization, predicting porosity and permeability, and lithology from well logs and seismic data. These examples demonstrate the diverse range of applications for machine learning in geology, geophysics, and petrophysics, helping researchers and professionals better understand and characterize subsurface geological formations.

The degree of bioturbation (vol. % of bioturbation) is mostly interpreted via visual estimation using the Bioturbation Index (BI) (Taylor and Goldring 1993). This can be a limitation that further augments the identification confusion between biogenic and diagenetic structures. This paper aims to efficiently improve the prediction of bioturbation in reservoir rocks via the SVC, k-NN, and LDA machine learning algorithms as has been done in some studies (e.g., Tarabulski and Reinhardt 2020; Zhang et al. 2021). It aims to serve as a framework for future bioturbation-related machine learning studies. This work combines the Support Vector Machine (SVM), k- Nearest Neighbour (k-NN), and Linear Discriminant Analysis (LDA) classification algorithms with wells data to effectively determine bioturbated zones in selected reservoir facies to reduce human error and provide more accurate outcomes.

Geological setting

Located on the western periphery of the Yellow Sea in northern Jiangsu province, eastern China, the Subei basin (Fig. 1) is characterized as a fault sag basin. Its geological history dates back to the Late Cretaceous period when it began as a rift and covers an estimated area of around 35,000 square kilometres (Song et al. 2010). The basin's formation can be divided into two primary rift phases, occurring between 83 and 54.9 million years ago, followed by another between 54.9 and 38 million years ago (Yang & Chen 2003; Chen 2010). The intervals of rift activity were separated by significant tectonic events known as the Wubao and Sanduo occurrences, associated with thermal subsidence, as documented by Liu et al. (2014). Liu et al. (2017) suggested the occurrence of the Wubao incident-induced faulting and division within the basin. In parallel, the Sanduo event resulted in notable uplift, subsequently causing erosion of the Oligocene and the strata beneath it. This process resulted in the forming of an angular unconformity observed between the Neogene and the formations positioned beneath it (Yi et al. 2003).

The paleocene funing formation (Ef)

The Paleocene Funing Formation's lowermost member (E_1f_1) displays a composition of 350 to 800 m of red beds (Fig. 2). These layers consist of intermixed brownish-red,

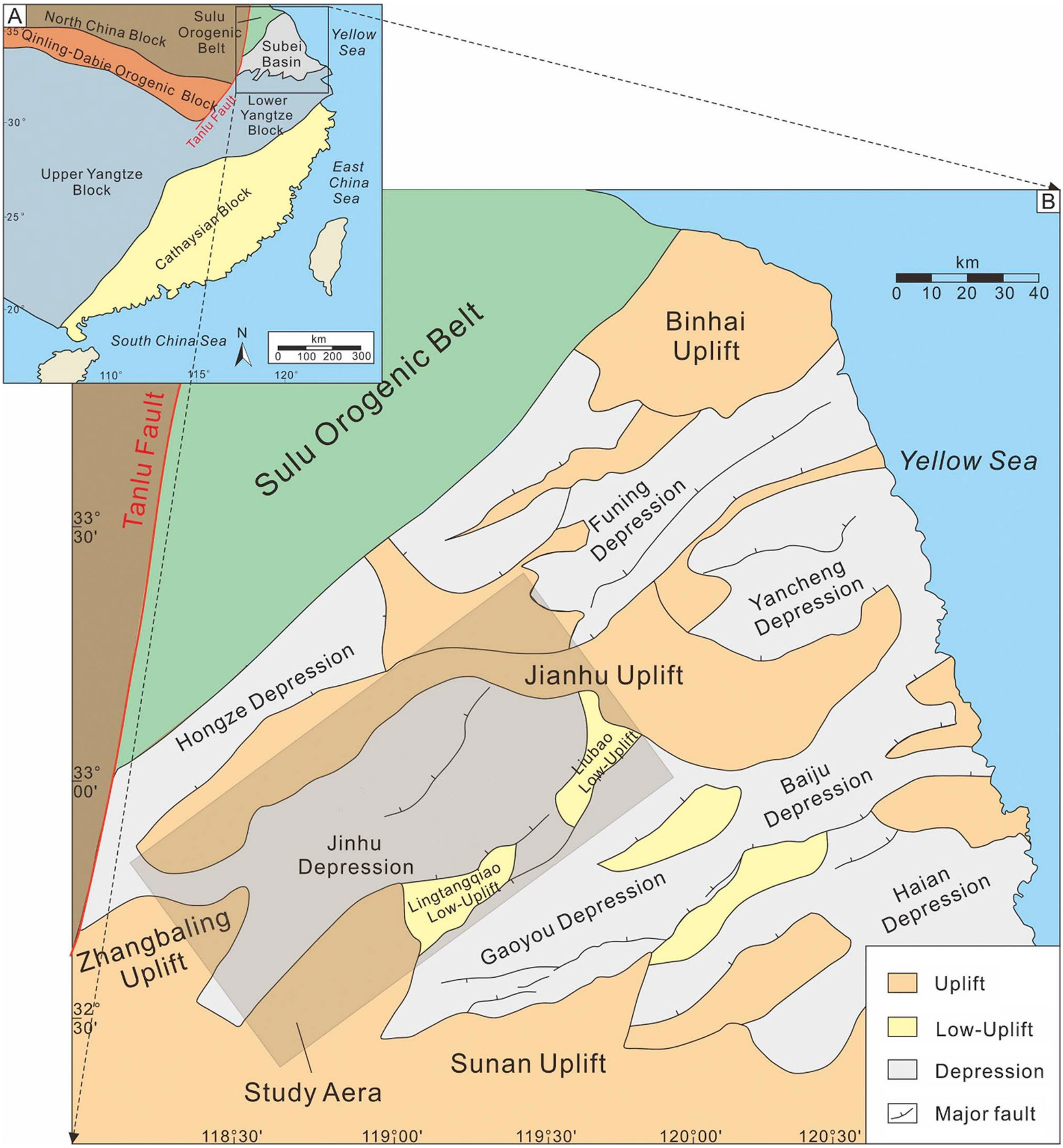

Fig. 1 **a** Description and maps detailing the general location of the Subei Basin general, positioned westerly to the Yellow Sea. **b** Delineation of different depressions within the Subei Basin, its surrounding tectonic features, and the specific examination of wells conducted in this research. Modified after Zhou et al. (2019)

very fine to fine-grained sandstones, siltstones, mudstones, and sporadic occurrences of greyish-green, very fine to fine-grained sandstones and siltstones (Zhang et al. 2006; Deng 2014). The Paleocene Funing Formation's second member (E_1f_2) comprises alternating layers of lacustrine carbonates and fine grey sandstones, spanning 70 to 110 m (Fig. 2), succeeded by a dark grey mudstone layer, 60 to 120 m thick (Liu et al. 2012; Luo et al. 2013; Shao et al. 2013). The third member (E_1f_3) includes layers of interbedded grey, very fine-grained mudstones and sandstones, spanning 200 to 300 m in thickness (Fig. 2). Zhang et al. (2006) suggested that this section is topped by the fourth member, a layer of dark grey mudstone that ranges from 300 to 400 m thick.

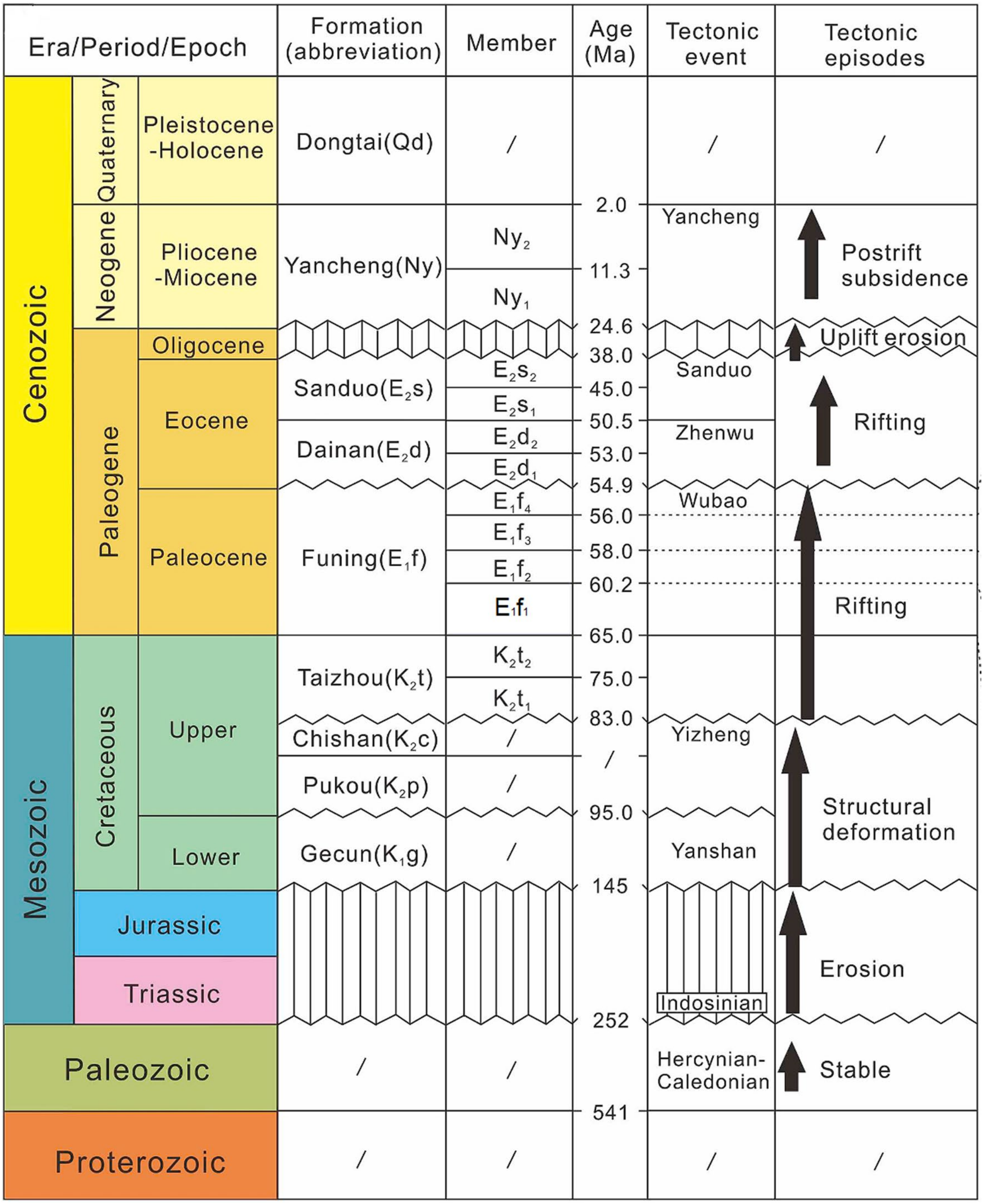

Fig. 2 In-depth analysis of the E1f1, E1f2, and E1f3 (65.0–56 Ma) involves a comprehensive study of their stratigraphy, facies, petroleum systems, tectonic occurrences, and evolutionary changes.Adapted from the work of Quaye et al. (2022)

Jinhu depression

The Jinhu Sag is situated southwest of the Subei Basin and is the largest within the basin, covering an area of approximately 5,500 square kilometres (Fig. 1B). Situated close to the northwest lies the Jianhu Uplift, while the Zhangbaling Uplift is positioned southwesterly, and the Sunan Uplift is found southeasterly. (Li et al. 2011; Liu et al. 2012; Shao et al. 2013). Within the Jinhu Depression, specific geological features are noteworthy, including the Liubao sandstone reservoirs and the Lingtangqiao Low-Uplift from the Paleocene Funing Formation (Li et al. 2011; Wang 2011). Dividing the formation (Fig. 2) into four clearly defined members has been accomplished using well logs and lithological studies (Liu et al. 2012).

Gaoyou depression

Positioned southerly to the Subei basin, the Gaoyou depression stretches about 2,670 square kilometres (Fig. 1B). Extending from east to west for over 100 kms and spanning about 30 kms from north to south, this depression takes the form of a half-graben. It is positioned with geographical separation to the east and south, demarcated by the Wubao Low and Tongyang uplifts.

The Jiangdu-Wubao fault zone delineates the southern and eastern perimeters of the depression, spanning over 140 kms and housing prominent faults like Wu 1, Wu 2, Zhen 1, and Zhen2. Within the Gaoyou depression, four discernible depocenters emerge Fanchuan, Liulu, Liuwushe, and Shaobo sub-basins (Liu et al. 2017).

In the western area, the Gaoyou Depression finds its boundaries marked by the Lingtangqiao Low Uplift, while the Tongyang Uplift characterizes its southern extent. During the Dainan-Yancheng period, significant movement occurred along two growth faults situated in the southern region of the Gaoyou depression. This movement resulted in the partitioning of the depression into three distinct segments as illustrated by Gu and Dai (2015): the South Fault-Terrace Belt, the Central Deep Sag, and the North Slope.

Bioturbation in the funing formation of the Subei basin

During the early Paleocene, the Subei basin was likely situated in a semiarid environment with seasonal rainfall patterns. This setting led to the development of diverse ichnofauna, including meniscal burrows, simple horizontal, vertical, or sub-vertical burrows, and plant roots/debris (Fig. 3G) along with their traces, which are characteristic of the Scoyenia or Skolithos ichnofacies in the Funing Formation. The Scoyenia ichnofacies predominantly feature horizontal meniscal burrows such as *Beaconites coronus* (Fig. 3A) *Taenidium satanassi* (Fig. 3B), and *Taenidium*

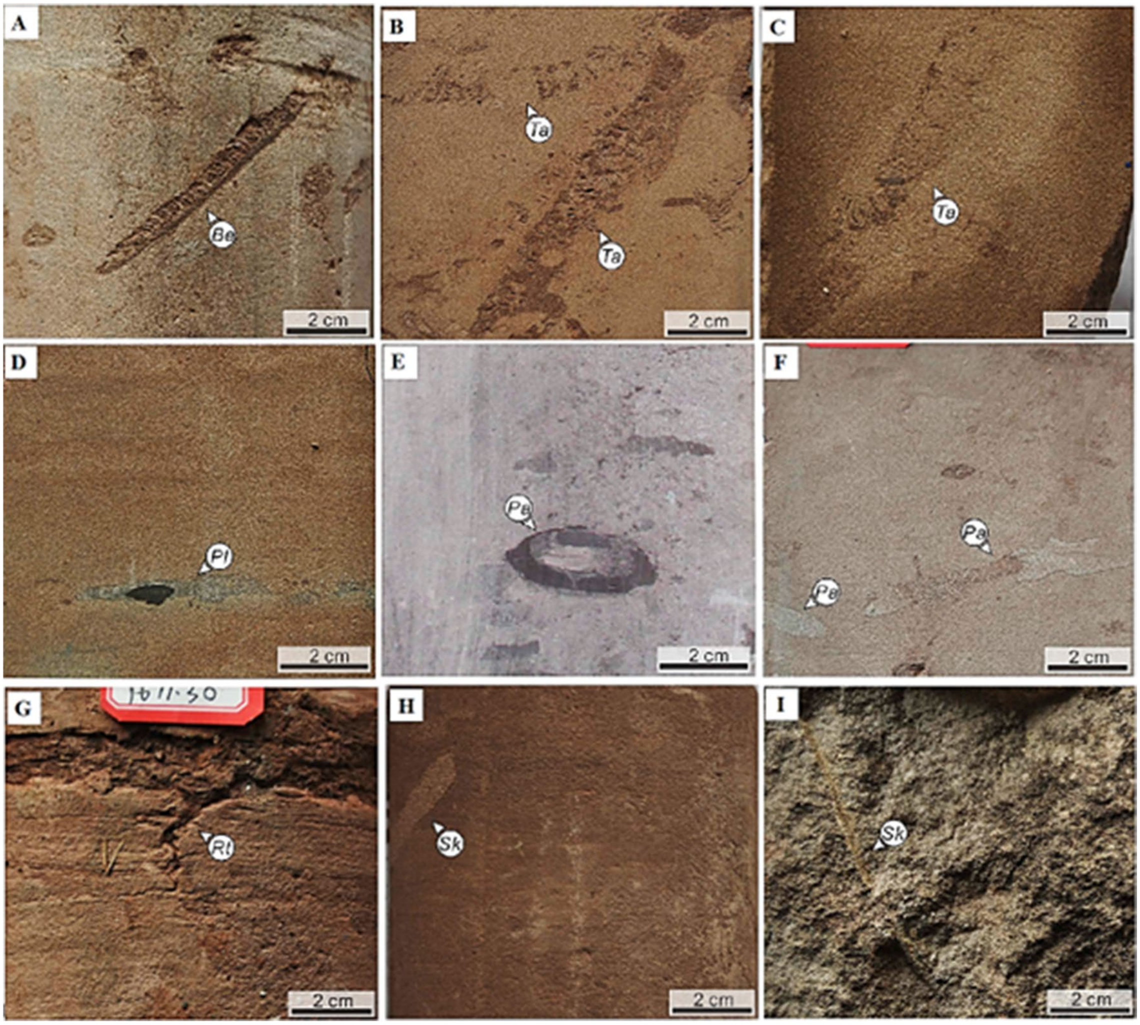

Fig. 3 **A** *Beaconites coronus* (Be); **B** *Taenidium satanassi* (Ta); **C** *Taenidium barretti* (Ta); **D** *Planolites isp.* (Pl); **E** *Palaeophycus heberti* (Pa); **F** *Palaeophycus tubularis* (Pa); **G** plant debris and/traces (Rt); **H** *Skolithos isp.* (Sk); **I** *Skolithos linearis* (Sk). Modified after Zhou et al. (2019)

barretti (Fig. 3C), along with simple horizontal cylindrical burrows like *Planolites* (Fig. 3D), *Palaeophycus heberti* (Fig. 3E), and *Palaeophycus tubularis* (Fig. 3F). The Skolithos ichnofacies include *Skolithos isp.* (Fig. 3H) and *Skolithos linearis* (Fig. 3I) (Zhou et al. 2019; Quaye et al. 2022, 2023). These ichnofacies are typically found in a mixture of clean, silty, and muddy substrates, indicative of multipurpose structures for feeding, dwelling, breeding, escape, and scavenging (Hubert and Dutcher 2010). They often exhibit very high bioturbation intensities ($4 \leq BI \leq 6$).

Methods

Data acquisition

This study used the Support Vector Machine (SVM), k- Nearest Neighbour (k-NN), and Linear Discriminant Analysis (LDA) classification algorithms combined with seventy-six (76) data points of nine (9) core samples (see Appendix) retrieved from five selected wells (see Fig. 1, B5; F12, H19, M7; and L5) in E_1f_2 of the Jinhu depression, and E_1f_1 and E_1f_3 of the Gaoyou Depression, respectively. These nine reservoir facies were mainly selected according to several indispensable factors. These are the outcrop area of facies, orientation and spacing of oilfield wellbores sectioned in relevant strata for ichnofauna search, and degree of bioturbation. The facies selected were also required to be free of fissures/fractures and any other defects that the results obtained would not accurately reflect their true values.

Table 1 presents a summary of core samples' properties that were considered as features for this work in the prediction of the BI. Key parameters were core dimensions, density, porosity, permeability, and the physical property's location.

Table 1 Core sample properties considered as features

Index	Parameters
1	Location
2	BI Vol
3	Particle Density
4	Volumetric Density
5	Porosity
6	Sample Diameter
7	Permeability
8	Pore Volume
9	Sample Length
10	Particle Volume
11	Dry Weight
12	Sample Volume

Data pre-processing and labelling

Data cleaning was performed to remove duplicates and handle missing values to achieve data quality for training. Then data labelling was achieved by assigning each row of core sample features with a bioturbation index classification.

The seventy-six data points from datasets of nine core samples were labelled with the appropriate BI that provides a degree of bioturbation from 0–6, hence seven classifications. Details of the data points are provided in the Appendix Tables 9, 10, 11, 12, 13, 14, 15, 16 and 17. Each data set was well-labelled based on expert advice and observation of cores.

Feature selection

Important features relevant to reservoir bioturbation were considered from the data set. These included seven relevant features extracted from the 12 features in the dataset, as listed in Table 2. SelectKBest is a common feature selection method that selects the best list of features based on statistical tests. The consideration of the ANOVA F-test is due to its suitability for high-dimensionality data sets. The comparison of variance between the BI classes to the variance with each group is also possible in the identification of features that possess a significant relationship with the BI index. Numerous investigators have hence considered the use of the ANOVA F test for feature selection in their supervised ML works (Shayestegan et al. 2024; Theng and Bhoyar 2024). Under the Python script, the ANOVA F-test (f_classif) is implemented via the SelectKBest statistical test used to score and rank features based on their relationship with the output variable.

The F-test statistics were calculated and defined in Eq. 1.

$$F - test = \frac{variance\ of\ features\ between\ the\ different\ classes}{variance\ of\ feature\ within\ each\ class} \tag{1}$$

The K feature with the highest f_score or f_value indicated a strong relationship between the features of the target. Feature scaling is applied to transform features in the dataset to a comparable scale and range to avoid the domination of features over others and to enable the models to perform better and converge quickly. The Minmax scaler is a feature scaling method that shrinks the features

Table 2 Summary of variables used for Linear Discriminant Analysis

Notation	Description
W	Transformation matrix
μ_i	The mean of the *i* th class
μ	The total mean
x_{ij}	The *i* th sample of *j* th class
S_{Bi}	Between-class variance
S_{wi}	Within-class variance

in a given dataset within the range of 0 to 1. Equation 2 shows the formula for Minmax scaler normalization.

$$Minmax = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2}$$

Data split

The dataset was divided randomly, allocating 80% for training purposes and reserving 20% for testing. The splits hence translate into 60 training and 16 test datasets, with their assigned labels. It is typical of a data split to have a higher percentage of training data set than the test set. A higher training set percentage will provide sufficient information for the models to be well-fitted to handle most of the feature combinations for adequate prediction of the BI (vol % of bioturbation). A data split of 80/20 was deemed suitable for the data set size in this study and also follows a more conservative approach according to Birba (2020). The test set is expected to be an unknown data set to asses model performance (Joseph and Vakayil 2022).

Model selection

In this work, three supervised classification ML methods are considered Support Vector Classification (SVC), K-Nearest Neighbour (K-NN), and Linear Discriminant Analysis (LDA). The models are trained using selected features as inputs to provide a robust model for BI (vol. % of bioturbation) prediction. The justification for the selection of these three classical supervised ML methods in the novel application of BI prediction is that each model has proven to handle small data sets with reasonable generalization. That is, the impact of the data sets on the performance of each model is considered insignificant especially after a thorough cross-validation of the models is performed (Beckmann et al. 2015; Nalepa and Kawulok 2019; Raikwal and Saxena 2012). It is well acknowledged that most recent ML algorithms exist such as the ensemble methods, however, for the sake of simplicity and novelty of application, it is worthy of consideration to commence with established and classical supervised ML algorithms to provide novel prediction of the BI. Furthermore, the ease of interpretability in the case of a small data set, in the use of the LDA, SVC and k-NN makes these a preferred model for this study. More advanced models can also be applied for the prediction of BI; however, numerous hyper-parameters need to be tuned to achieve an optimal model. Although the selected models come with some limitations associated with the choice of kernel function (for SVC), choice of k neighbour (in the case of k-NN) and limitations to multi-dimensionality given linear boundary (in the case of LDA), the type of classification problem presented in work affords their use as the advantages outweigh the limitations.

Linear Discriminant Analysis (LDA)

The LDA serves as a supervised machine learning algorithm designed to execute classification tasks effectively. Additionally, it's adept at addressing dimensionality reduction challenges, eliminating redundant and interdependent features, and transforming high-dimensional features into a more concise low-dimensional space (Tharwat et al. 2017). All classes are assumed to be linearly separable, and hyperplanes within the feature space are created to differentiate between classes. (Vaibhaw and Pattnaik 2020). Hyperplanes are created based on two criteria: first, maximizing the separation between the means of distinct classes, known as the between-class variance (S_{Bi})) as shown in Eq. 3; secondly, minimizing the distance between the class means and their respective samples, termed the within-class variance (S_{Wi})) as represented in Eq. 4. The various variables used in this paper are described in Table 2.

$$S_{Bi} = W^T(\mu_i - \mu)(\mu_i - \mu)^T W \tag{3}$$

$$S_{wi} = W^T(x_{ij} - \mu_j)(x_{ij} - \mu_j)^T W \tag{4}$$

Support Vector Classification (SVC)

The Support Vector Classifier (SVC) represents a supervised machine-learning algorithm specifically applied to address multi-classification challenges. It operates in a fashion similar to LDA. The SVM creates decision boundaries between classes that help predict labels from feature vectors (Huang et al. 2018). Decision boundaries are called hyperplanes. The number of dimensions in the data dictates the configuration of hyperplanes. Through a set of constraints, Support Vector Machines (SVM) engage in an optimization process to ascertain optimal hyperplanes that maximize the margin between distinct classes. This margin denotes the space between the hyperplane and the support vector, which represents the closest data point from each class. Equation 5 defines the hyperplane equation, where 'w' signifies the weight, and 'b' stands for the bias.

$$f = w.x + b = 0 \tag{5}$$

The objective is to identify a hyperplane that optimizes margins while minimizing classification errors. There is a need to optimize the quadratic function with linear constraints defined in Eq. 6 and subjected to Eq. 7, where y_i denotes training data class label, point x_i.

$$minimize = \frac{1}{2}|w||w| \tag{6}$$

$$subjected : y_i(w.x_i + b) \geq 1 for all i \tag{7}$$

The optimal hyperplanes can separate data points defined by the decision rule in Eq. 8.

$$\begin{cases} +1, ifw.x + b \geq 0 \\ -1, ifw.x + b \leq 0 \end{cases} \tag{8}$$

K-Nearest Neighbour

The K-Nearest Neighbour (K-NN) algorithm enjoys widespread use in classification due to its simplicity in computation and interpretation (Moldagulova and Sulaiman 2017). The K-NN algorithm relies on distance metrics, evaluating the similarity between two points based on their distance. The commonly utilized distance metric in scikit-learning is the Euclidean distance function, determining the distance between points X and Y. This Euclidean distance is mathematically defined in Eq. 9.

$$|XY| = \sqrt{\left(x_1 - y_1\right)^2 + \left(x_2 - y_2\right)^2} \tag{9}$$

The k value is the number of numbers used to make a prediction. Choosing k is important because it significantly affects the algorithm's performance. The proximity of the nearest k values was evaluated and sorted based on their closeness. The K-NN algorithm predicts the class of a new data point based on the majority class of the K most similar data points.

Model training and hyper parameter tuning

The model training entails the use of an 80% data set (66 training data points) to provide a multidimensional fit of decision boundaries for each model. To avoid the problems of underfitting and overfitting, optimization of the hyperparameters of the SVC and K-NN models was performed to ensure that models minimize the loss function. For the LDA, no optimization was required since there are no hyperparameters in the algorithm framework. For instance, the SVC model has the C and gamma and kernel parameters which require optimization or tuning while the K-NN model has the K-nearest neighbour’s value to optimize decision boundaries before further cross-validation can be performed. The C parameter typically represents the regularization parameter, playing a crucial role in determining the tradeoff between bias and variance. A smaller C provides for a larger margin separating the hyperplane and a larger C parameter leads to a smaller margin separating the hyperplane that provides the decision as to the varied BI (vol.% of bioturbation) classifications. The kernel parameter defines the type of decision boundary such as linear, polynomial or sigmoid (radial basis function) in nature. The Gamma parameter defines the degree of influence of a single training data on the decision boundary. Hence, high values of gamma depict a closer degree of influence and vice versa for a low-value situation. For the K-NN as a classifier, the kth nearest neighbour classifies new data to be within a class based on the proximity of the new data to k number of classes labelled within a defined distance termed the neighbourhood. If K is low, the algorithm tends to capture local patterns within the data but it is short of handling noise or outliers in the data. A high K value may provide a smother decision boundary but may not capture local variations in the data. Therefore, the need for k-parameter tuning and optimization.

The grid search method is applied in the search for all hyper-parameter combinations that are considered for the optimal multidimensional grid. Hyperparameter optimization will aid adequate bias-variance tradeoff that will provide model robustness in the prediction of unseen test datasets.

Model cross-validation

It is important to attain model stability and generalization which infers that the classification models are independent of the training data set selection. To ensure consistent accuracy and model reliability, cross-validation was executed on the training datasets. This process evaluated how well the machine learning models performed on unseen data, aiming for generalization and stability. The K-fold cross-validation was applied with a value of k set to 5. The training data set was further split into 80% sub-training and 20% cross-validation sets (illustrated in Fig. 4). In the fivefold, the models were trained and evaluated, such that, training and cross-validation were performed five times with each period using a different form of sub-training

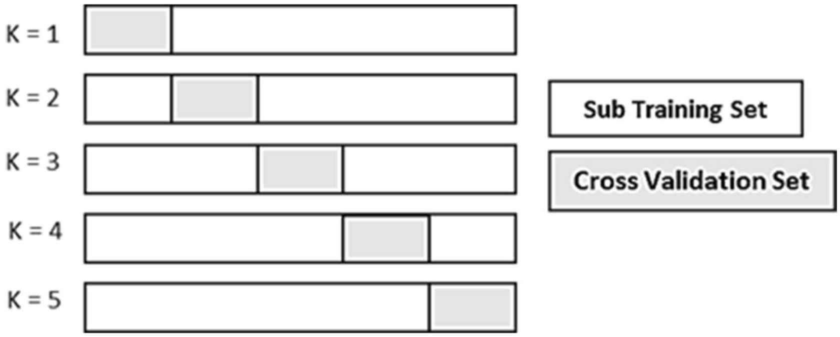

Fig. 4 5-Fold cross validation step

sets and cross-validation sets. Results from the cross-validation will provide insight into the data dependence and stability of classification models. A cross-validation step is also required to prevent model underfitting and overfitting issues.

Model performance evaluation

Given that the issue addressed in this study pertains to multiclass classification, a confusion matrix was considered for the model performance evaluation. A detailed treatment of the confusion matrix for multiclass model assessment has been explored by some researchers (Delgado and Núñez-González 2019; Mathur and Foody 2008). Key metrics from the confusion matrix include average error, F1 score, Recall, precision and average accuracy. Performance metrics considered for this work will be based on average accuracy and average error. Overall accuracy based on the training set, cross-validation or test sets was determined as the ratio of accurately predicted BI (vol.% of bioturbation) to the total actual BI classifications. The model performance average error was evaluated based on loss of the target, defined as, 1-average accuracy. Precision refers to the ratio of the correctly predicted positive observations referred to as true positives to the total predicted positive observations.

$$Precision = \frac{True\ positives}{True\ positives + False\ Positives}.100\% \tag{10}$$

Recall is defined as the ratio of True positives to the sum of the true positives and False negative predictions

$$Recall = \frac{True\ positives}{True\ positives + False\ Negatives}.100\% \tag{11}$$

The F1 score is the weighted average of the Precision and Recall such that

$$F1score = 2.\frac{Precision\ x\ Recall}{Pricision + Recall}.100\% \tag{12}$$

Figure 5 summarizes the key steps of processes utilized via machine learning.

Results and discussion

Feature selection

In this work, a total of twelve features (as summarized in Table 3) were considered to be screened based on relevance and significance to changes in bioturbation prediction. For feature selection, based on f-scores and p-values, 7 features indexed as 1,2,3,4,5,6,7, were selected to be

Fig. 5 Summary flow-chart process of machine learning

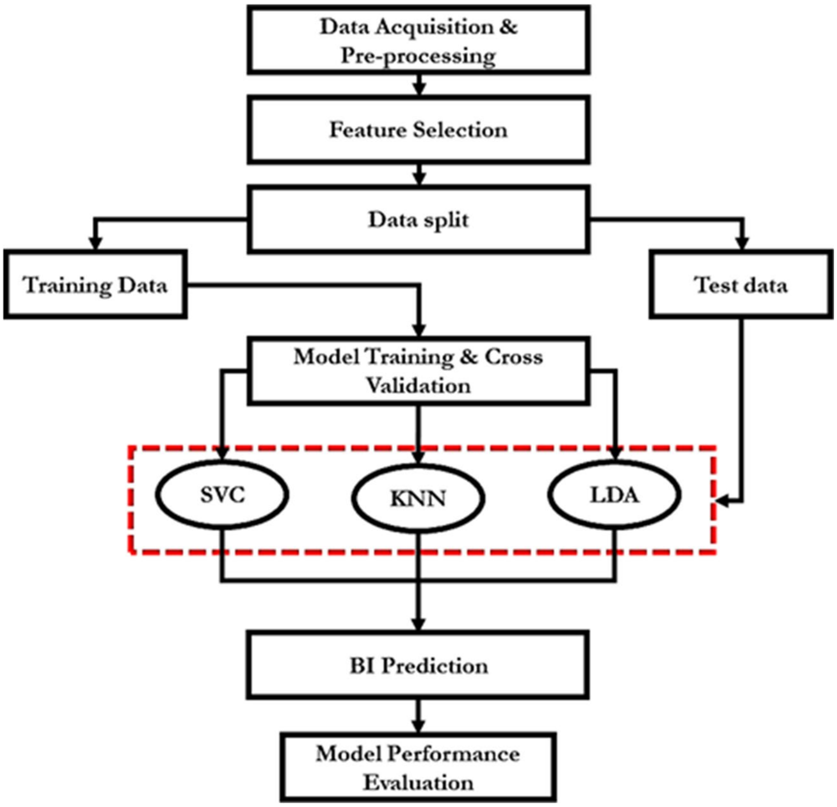
Data Acquisition & Pre-processing
Feature Selection
Data split
Training Data
Test data
Model Training & Cross Validation
SVC
KNN
LDA
BI Prediction
Model Performance Evaluation

Table 3 Feature selection

Index	features	f_score	*P*_values	Decision
1	Location	inf	0	TRUE
2	BI Vol	23742.85	0	TRUE
3	Particle Density	4.87334	0.00157	TRUE
4	Volumetric Density	4.28574	0.00366	TRUE
5	Porosity	1.94269	0.11275	TRUE
6	Sample Diameter	1.19159	0.32203	TRUE
7	Permeability	1.05612	0.38469	TRUE
8	Pore Volume	0.86715	0.48803	FALSE
9	Sample Length	0.58844	0.67206	FALSE
10	Particle Volume	0.58601	0.67379	FALSE
11	Dry Weight	0.47752	0.7521	FALSE
12	Sample Volume	0.38021	0.82206	FALSE

used as inputs for each of the supervised ML models. The remaining 5 features indexed as 8, 9, 10, 11, and 12 were rejected since low f-scores and high p-values were obtained. Features with p-values greater than 0.45 were rejected since increases in p-value represent a higher probability that the respective feature adds little or no information to predicting bioturbation.

Put differently, it means that features with high p-values possess a high likelihood of changing the BI (vol.% of bioturbation) by chance and by a redefined correlation with the BI (vol.% of bioturbation). The rejected features with high p-values also corresponded with increases in f- f-scores.

Model training performance

Training of the classifiers was performed based on the selected features (see Table 3) with hyperparameter tuning of SVC and K-NN models using the training set. In Fig. 6, the results of the effects of the SVC hyperparameters, Gamma, Kernel and C on the average training accuracy of the bioturbation index (vol % of bioturbation) are shown. It can be deduced from the figure, that a general increase in training accuracy is observed for increases in the regularization parameter C for varied values of the gamma and kernel functions considered. This is simply because increasing the C parameter leads to a reduced regularization and a more complex decision boundary, hence a smaller boundary margin separating the bioturbation index classes such that the hyperplane accurately classifies the new data. Although very high C values (small regularization) can lead to overfitting of the training set, a balance between generalization and complexity is key to achieving the optimal classifier model.

For the linear kernel function, under each C value that ranges from 1–1000, increases in the gamma parameter from 0.01 to 1.00 showed little or no effect on the training accuracy. In other words, the variation in similarity radius for each class of bioturbation index does not affect the training fit of the SVC model. For example, at C = 1, under the linear kernel, increases in the gamma from 0.01 to 1.00, showed similar training accuracies around 62%. A similar training accuracy trend can be observed for further increases in C = 10, 100, and 1000

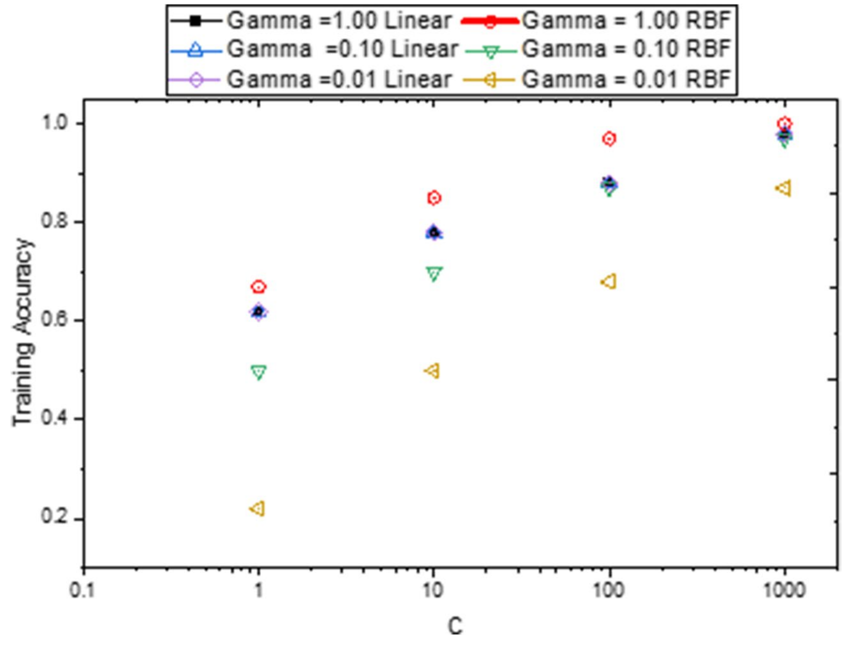

Fig. 6 Effects of C, Kernel and gamma on the training performance

under a linear kernel to be around 78.4%, 84.2% and 96.1% respectively. The insensitivity of the SVC prediction to the gamma parameter is mainly due to the simpler linear function that defines the decision boundaries for the classification of the BI (vol.% of bioturbation) for the training data set considered.

Conversely, under the RBF kernel, which is a more complex function, the effect of the gamma parameter is evident given the clear distinction in training accuracies for a given C parameter (as shown in Fig. 6). The more complex sigmoid nature of the RBF allows for the effect of the gamma parameter that represents changes in the similarity radii of each BI class to be significant, hence leading to increases in fitting performances of the SVC model, that amount to training accuracies as high as 98% (for C = 1000 and gamma = 1.00).

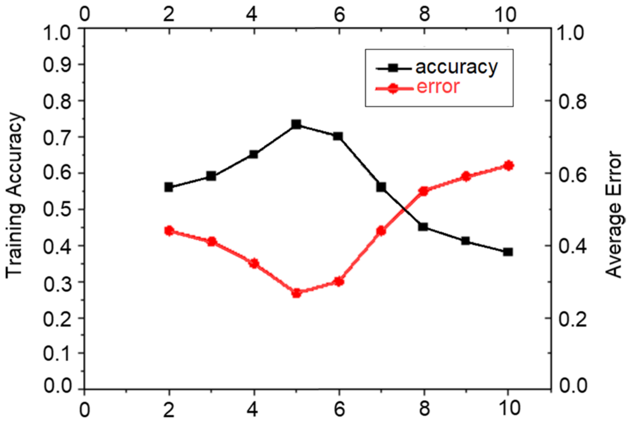

Fig. 7 Effect of the kth nearest neighbour on the training accuracy of the K-NN model

Overall, the optimal hyperparameters of the SVC model selected for this work are C = 1000, and Gamma = 0.10 based on the linear kernel function. The selection of these hyperparameters can be justified to obtain an SVC model with an adequate bias-variance tradeoff. The results of the RBF kernel function also show less stability in training accuracy compared to the linear function case where there is little or no effect of the gamma parameter on the training accuracy.

Figure 7 presents the results of the effect of the k value on the training accuracy and corresponding error rates for the K-NN model. A variation of k values from 1–10 indicates a maximum training accuracy at k = 5 from the gird search results which corresponds to a training accuracy of 73.28%. For increases in k values (less than k = 5), decision boundaries of the K-NN model tend to be more complex and hence sensitive to outliers in the training data set since it relies heavily on the nearest neighbour classification. For K values greater than 5 the K-NN model tends to decrease in training accuracy (increase in error rate) as an underfitting results from the simpler decision boundaries that misclassify the training data. K = 5 is therefore selected as the optimal hyperparameter for the K-NN model.

Table 4 Confusion Matrix Training performance of hyperparameter tuned SCV model

SVC Model								
Actual Bioturbation Index	Predicted Bioturbation Index							
	Class 0	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Total
Class 0	13	0	0	0	0	0	0	13
	100	0	0	0	0	0	0	100
Class 1	0	8	0	0	0	0	0	8
	0	100	0	0	0	0	0	100
Class 2	0	0	6	0	0	0	0	6
	0	0	100	0	0	0	0	100
Class 3	0	0	1	6	0	0	0	7
	0	0	14.29	85.71	0	0	0	100
Class 4	0	0	0	0	8	0	0	8
	0	0	0	0	100	0	0	100
Class 5	0	0	0	0	1	7	0	8
	0	0	0	0	12.5	87.5	0	100
Class 6	0	0	0	0	0	0	10	10
	0	0	0	0	0	0	100	100
ERROR	0	0	0	14.29	0	12.5	0	**3.83**
ACCURACY	100	100	100	85.71	100	87.5	100	**96.17**

Table 5 Confusion Matrix showing training performance of hyperparameter optimized KNN model

KNN model								
Actual Bioturbation Index	Predicted Bioturbation Index							
	Class 0	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Total
Class 0	13	0	0	0	0	0	0	13
	100	0	0	0	0	0	0	100
Class 1	0	6	2	0	0	0	0	8
	0	75	25	0	0	0	0	100
Class 2	1	0	5	0	0	0	0	6
	16.67	0	83.33	0	0	0	0	100
Class 3	0	0	2	4	1	0	0	7
	0	0	28.57	57.14	14.29	0	0	100
Class 4	0	0	0	0	4	3	1	8
	0	0	0	0	50	37.5	12.5	100
Class 5	0	0	0	0	1	5	2	8
	0	0	0	0	12.5	62.5	25	100
Class 6	0	0	0	0	2	2	6	10
	0	0	0	0	20	20	60	100
ERROR	0	0	16.67	42.86	50	37.5	40	**26.72**
ACCURACY	100	75	83.33	57.14	50	62.5	60	**73.28**

Training results of optimized classifiers

This section presents the training results of the optimized classifiers (SVC and K-NN) and LDA models in the form of the confusion matrices presented in Tables 4, 5 and 6. The confusion matrix presented in Table 4 shows the training results of the optimized SVC in the prediction of the Bioturbation for the training data set (60 datasets). For the SVC, most classes of bioturbation were excellently classified and predicted relative to the actual BI (vol.% of bioturbation) classification. However, classes 3 and 5 were adequately predicted but the false positive predictions for classes 2 and 4

Table 6 Confusion matrix showing training performance LDA model

LDA Model								
ACTUAL BIOTURBATION INDEX	Predicted Bioturbation Index							
	Class 0	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Total
Class 0	13	0	0	0	0	0	0	13
	100	0	0	0	0	0	0	100
Class 1	1	7	0	0	0	0	0	8
	12.5	87.5	0	0	0	0	0	100
Class 2	0	0	5	1	0	0	0	6
	0	0	83.33	16.67	0	0	0	100
Class 3	0	0	0	6	0	1	0	7
	0	0	0	85.71	0	14.29	0	100
Class 4	0	0	0	0	3	4	1	8
	0	0	0	0	37.5	50	12.5	100
Class 5	0	0	0	0	3	3	2	8
	0	0	0	0	37.5	37.5	25	100
Class 6	0	0	0	0	3	3	4	10
	0	0	0	0	30	30	40	100
ERROR	0	12.5	16.67	14.29	62.5	62.5	60	**32.64**
ACCURACY	100	87.5	83.33	85.71	37.5	37.5	40	**67.36**

Table 7 Other performance metrics for all models in this study

Metric	Class 0	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Average
SVC								
Precision	100.00%	100.00%	100.00%	85.71%	100.00%	87.50%	100.00%	96.17%
Recall	100.00%	100.00%	100.00%	85.71%	100.00%	87.50%	100.00%	96.17%
F1 Score	100.00%	100.00%	100.00%	85.71%	100.00%	87.50%	100.00%	96.17%
KNN								
Precision	100.00%	75.00%	83.33%	57.14%	50.00%	62.50%	60.00%	69.71%
Recall	100.00%	75.00%	83.33%	57.14%	50.00%	62.50%	60.00%	69.71%
F1 Score	100.00%	75.00%	83.33%	57.14%	50.00%	62.50%	60.00%	69.71%
LDA								
Precision	100.00%	87.50%	83.33%	85.71%	50.00%	37.50%	40.00%	69.15%
Recall	100.00%	87.50%	83.33%	85.71%	37.50%	37.50%	40.00%	67.36%
F1 Score	100.00%	87.50%	83.33%	85.71%	42.86%	37.50%	40.00%	68.13%

respectively led to errors of 14.29% (for class 3) and 12.50% (for class 5). An overall training accuracy of 96.17% was obtained (3.83% error) under the optimized SVC.

In comparison with the K-NN method (Table 5), an overall training accuracy of 73.28% corresponding with an average error of prediction of BI of 26.72% is obtained. In the K-NN model, there were some misclassifications relative to the actual labels of BI on the data sets considered given the choice of k = 5.

As presented in Table 6, the LDA method provides training overall accuracy of 67.36% and an error of 32.64%. Under the LDA method, a linear combination of the features is expected to predict the BI (vol.% of bioturbation), hence the Significant error obtained could be related to an underfitting problem which may require further cross-validation.

The confusion matrix shows the classification performance across seven classes. The precision, recall, and F1 score are perfect (1.00) for Classes 0, 1, 2, 4, and 6, indicating flawless classification with no false positives or false negatives.

Class 3 has a lower performance with a precision and recall of 0.86, reflected in its F1 score of 0.86. This reduction is due to one instance of Class 2 being misclassified as Class 3 (false positive) and one instance of Class 3 being misclassified as Class 2 (false negative). This suggests some confusion between these two classes.

Class 5 also shows slightly reduced performance, with a precision and recall of 0.88 and an F1 score of 0.88. This is due to one instance of Class 4 being misclassified as Class 5 and one instance of Class 5 being misclassified as Class 4, indicating a minor overlap between these classes.

Overall, the model demonstrates high accuracy, correctly classifying the vast majority of instances. The slight misclassifications in Classes 3 and 5 suggest areas for further refinement, potentially focusing on distinguishing features between the closely related classes.

Other performance metrics

The results of other performance metrics such as the precision-recall and F1 score of each model are presented in Tables 7 and 8. Training results based on other metrics such as the precision, recall and F1 score for each model are presented in Table 7. The results indicate that the SVC model outperforms other models with an average precision, recall and score of 96.17%.

More so, on a class basis, the SVC still performs best with the least training classification performances in class 3 at 85.71%. This indicates that the SVC model could be effective in accurately predicting the bioturbation index across the different classes with minimal error.

The KNN model also shows relatively lesser training performance than the SVC model, although with perfect recall, precision and F1 scores for class 0, other classes were moderately classified correctly with the least precision-recall and f1 score to be 50 per cent at class 4 predictions. Overall, the F1 score of the KNN model training was 69.71%. These performances suggest that the KNN model has some overlapping class boundaries given imbalanced

Table 8 Test performance results on BI prediction

Bioturbation Index	Test Data		
	SVC	KNN	LDA
Class 0	100	100	100
Class 1	100	100	50
Class 2	50	100	50
Class 3	100	66.7	66.7
Class 4	100	100	100
Class 5	100	66.7	66.7
Class 6	100	100	100
Average Accuracy	**92.86**	**90.48**	**76.2**

class distributions relative to the SVM model. The LDA model that provides linear class boundaries also performs best for classes 1,2,3 at F1 scores above 80%. However, low training classifications especially for classes 4, 5 and 6 were as low as 37.50%. The overall precision, recall and F1 scores of 69.15%, 67.36% and 68.13% respectively indicate that the LDA model given its linear boundaries tends to misinterpret overlapping classes.

Cross-validation of models

To establish the generalization of each model performance, a cross-validation using the fivefold stratification of the training data set was performed. Figure 8(a) below shows the error as a result of variation in training sets used for each model. The stability of each model can be inferred from the variation in average error for increases in k-folds from 1 to 5. For instance, for the SVC model variation in error is between 1% (at a K-fold of 1) and 6.25% (at a K-fold of 4). The undulating behaviour within the stated average errors is indicative of a stable model and independence of the training data set used. For the K-NN model, similar stability of training performance can be inferred, albeit the significant losses in prediction. The LDA model shows a relative increase in loss with an increase in K folding of the training set combined with the highest average errors in the prediction of BI (vol.% of bioturbation). The LDA model is, therefore, considered the worst-performing model in this work. The results of average cross-validation accuracy of the prediction of BI (vol % of bioturbation) for all models are depicted in Fig. 8(b). The results of the average cross-validation accuracies indicate that the optimized SVC is the best-performing classifier followed by the optimized K-NN classifiers and the worst perming to be the LDA classifier.

Test performance

Table 7 showcases the test performance outcomes for BI (vol. % of bioturbation) prediction across the SVC, K-NN, and LDA classifiers. The support vector classifier outperforms other classifiers in the prediction of BI (vol.% of bioturbation) overall, which results in an average accuracy of 92.86%. Although a poor prediction of the class 2 BI (vol % of bioturbation), as 50% cannot be ignored more data is recommended to improve the results of classification. The KNN classifier performs at an average accuracy of 90.48% which is acceptable. Classes 3 and 5 were predicted at 66.7% accuracy as a result of the decision boundary defined by the K-NN model. The linear discriminant model remains the worst-performing classifier in this work with an average accuracy of 76.2% with only classes 0, 4, and 6 predicted at 100% accuracies for the test data set considered.

In comparison with other works such as that of Timmer et al. (2021) that considered deep learning methods (deep convolutional neural networks (DCNNs)) using images as input and arrived at an 88% accuracy of prediction of bioturbation. More so, Ayranci et al. (2021) also obtained an accuracy of 70% when they used a neural network algorithm combined with a high number of input images to detect the Bioturbation Index (vol % of bioturbation). In this work, results indicate an improved prediction of the BI especially with the SVC when given the relatively small sample space and expert labelling of core samples with BI (vol.% of bioturbation) within the respective classes of identification. The improved performance of the SVC compared to neural networks in a classification problem is due to the capability of the kernel function to augment feature numbers onto which the intrinsic data properties are extracted.

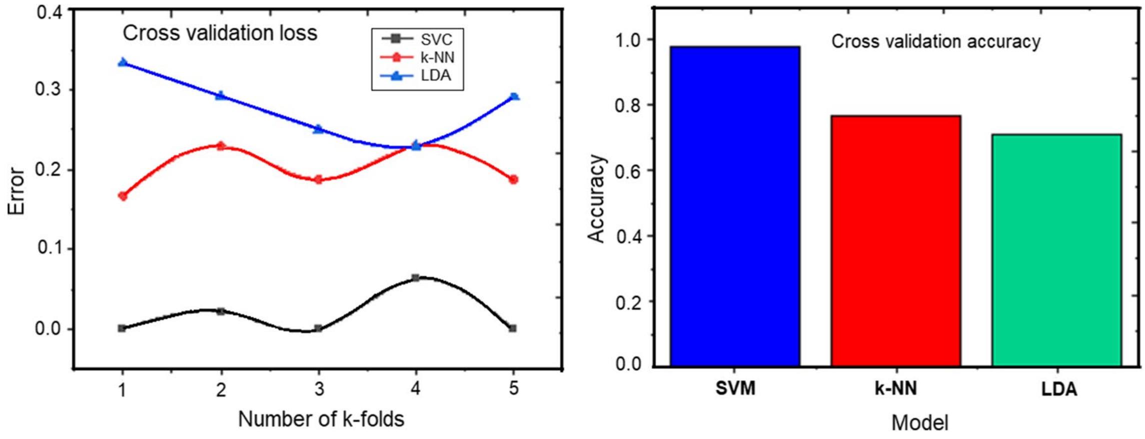

Fig. 8 Cross-validation results of the fivefold cross-validation on (**a**) error rates for SVC, K-NN and LDA classifiers and (**b**) cross-validation accuracy

Models' prediction limitations on bioturbated reservoir facies

Reservoir-scale advantages of studying bioturbation mainly focus on a better understanding of mud-dominated sedimentary structures that allow improved predictions for rock properties (e.g., Buatois and Mángano 2011; Gingras et al. 2001). It is also of special interest in hydrocarbon reservoirs since it modifies porosity and permeability. Yet, they are also fraught with a range of major limitations and difficulties (Gibling and Bird 1994). Pros and cons of applying Machine Learning (ML) models to predict bioturbation in reservoir conditions They are particularly good at churning through vast amounts of data, uncovering patterns and links that traditional methods may not identify. Reservoir conditions such as fractures, fissures, and diagenetic processes can substantially impact the accuracy and reliability of these ML models. Furthermore, fractures and fissures lead to complex flow pathways within the reservoir, hence complicating bioturbation signal interpretation with a potential reduction in the performance of ML predictions (Oliver et al. 2008; Tarabulski and Reinhardt 2020). Baniak et al. (2013) propose that bioturbated reservoir facies may be permeable with areas of low porosity acting as semi-sealed, intra-stratum micro-fracture systems. Such housings create fluid flow focussed on these high-permeability paths, which may be decoupled by structural features and pass deposition away from (bioturbated matrix), complicating predictions. Due to diagenetic processes such as cementation, dissolution and recrystallization at the micro-scale the primary sedimentary structures or bioturbation features could be modified which represents complications in interpretation because these elements may affect final model surrogates for hydraulic properties. An insight into these interactions is important for improving the predictive capabilities of ichnological models in a wide range of reservoir conditions (Worden and Burley 2003; Buatois and Mángano 2007).

Conclusion

This work considers a unique set of inputs that includes the key dimensions of the core samples and the volume of bioturbation in the sample is considered for the prediction of BI via SVC, K-NN and LDA algorithms. These classifiers provide decision boundaries that aid the prediction of the multi-classification of Bioturbation in the form of the Bioturbation index. 76 data sets of core samples retrieved from existing wells in the Subei Basin, China. Key steps in machine learning performed in this work include data preprocessing, feature selection, model training cross-validation, and testing. Seven (7) selected features from the core data were used as inputs to build each classifier to predict bioturbation.

A training-test data split of 80/20 was adequate for the study. Training of the SVC and K-NN models considered hyperparameter optimization and cross-validation of all models before a model evaluation using the test data set. Based on grid search, the hyperparameters of the SVC and K-NN models were selected based on adequate bias-variance tradeoff considerations. The training and test results indicate that the optimized SVC was the best classifier followed by the k-NN classifier and then the LDA classifier which was the worst-performing classifier.

The results also show that hyperparameter optimization is critical for desired model performances. The novelty of this work was evident in the application of core data that comprised rock properties and BI parameters as selected features for training each classifier to predict bioturbation compared to other works that considered images of core samples as features for bioturbation prediction. We recommend adaptive unsupervised ML classifiers to predict bioturbation in future works.

Appendix

Tables 9, 10, and 11 data are extracted and modified after Quaye et al. (2022).

Tables 12 and 13 data are extracted and modified after Quaye et al. (2019)

Tables 14, 15, 16, and 17 data are extracted and modified after Quaye et al. (2023)

Table 9 Physical properties of sample 1 retrieved from well M7 in the E_1f_1

Sample Length	Sample Diameter	Sample Volume	Dry Weight	Particle Volume	Pore Volume	Porosity	Particle Density	Volumetric Density	Permeability	Location	Bioturbation Index	BI Vol
(cm)	(cm)	(cm3)	(g)	(cm3)	(cm3)	(%)	(g/cm3)	(g/cm3)	(md)	(B/M)	BI	(%)
2.146	2.528	10.766	26.84	10.06	0.7015	6.5122	2.49	2.67	0.016	Burrow	5	96
4.246	2.526	21.267	53.86	20.52	0.7446	3.4995	2.53	2.62	0.0156	Burrow	1	4
2.394	2.524	11.972	30.28	11.52	0.4557	3.8041	2.53	2.63	0.035	Burrow	6	100
2.384	2.538	12.055	30.57	11.12	0.9343	7.7463	2.54	2.75	0.0246	Burrow	5	96
4.142	2.536	20.911	52.22	19.28	1.6342	7.8109	2.5	2.71	0.0241	Matrix	0	0
1.946	2.532	9.794	24.35	8.8	0.9977	10.1828	2.49	2.77	0.0123	Burrow	3	60
2.218	2.518	11.039	28.07	10.15	0.8923	8.0784	2.54	2.77	0.0282	Burrow	4	90
4.14	2.528	20.769	52.71	19.47	1.2974	6.2435	2.54	2.71	0.0214	Burrow	5	96
2.446	2.534	12.329	30.76	11.71	0.6172	5.0032	2.49	2.63	0.0347	Burrow	2	30
2.206	2.528	11.067	27.61	10.52	0.5508	4.974	2.49	2.63	0.036	Burrow	6	100
4.24	2.534	21.372	54.41	19.96	1.4081	6.5851	2.55	2.73	0.021	Burrow	4	90
2.1	2.528	10.535	26.11	9.56	0.9717	9.219	2.48	2.73	0.0358	Burrow	3	55
2.232	2.532	11.233	27.76	9.96	1.2723	10.03	2.47	2.79	0.0091	Burrow	4	90
4.552	2.522	22.728	57.16	21.17	1.562	6.869	2.51	2.7	0.02	Matrix	0	0
2.308	2.536	11.652	28.72	10.68	0.9736	8.3514	2.46	2.69	0.0119	Burrow	2	25
1.95	2.528	9.783	24.59	9.27	0.5152	5.2636	2.51	2.65	0.0091	Matrix	0	0
3.53	2.538	17.85	45.24	16.87	0.9766	5.4682	2.53	2.68	0.0133	Burrow	1	2
2.388	2.524	11.942	29.73	11.24	0.7022	5.8771	2.49	2.65	0.0561	Burrow	3	60

Table 10 Physical properties of sample 2 retrieved from well F12 in the E_1f_1

Sample Length	Sample Diameter	Sample Volume	Dry Weight	Particle Volume	Pore Volume	Porosity	Particle Density	Volumetric Density	Permeability	Location	Bioturbation Index	BI Vol
(cm)	(cm)	(cm3)	(g)	(cm3)	(cm3)	(%)	(g/cm3)	(g/cm3)	(md)	(B/M)	BI	(%)
1.716	2.518	8.541	20.49	8.01	0.5324	6.2307	2.4	2.56	0.3674	Burrow	2	20
3.708	2.528	18.602	44.53	17.37	1.2316	6.6175	2.39	2.56	0.455	Burrow	4	90
1.676	2.514	8.315	19.75	7.56	0.7533	9.0549	2.38	2.61	0.9349	Burrow	6	100
1.334	2.534	6.724	15.66	6.6	0.1201	1.7847	2.33	2.37	0.4474	Matrix	0	0
3.908	2.52	19.482	47.22	17.61	1.8669	9.578	2.42	2.68	0.2947	Burrow	4	90
1.598	2.504	7.865	18.82	7.38	0.4836	6.145	2.39	2.55	0.2886	Matrix	0	0
1.725	2.511	8.543	20.55	8.68	0.7266	8.505	2.41	2.37	0.501	Burrow	3	35
1.927	2.501	9.468	22.65	9.05	0.7311	7.7219	2.39	2.5	0.5309	Burrow	5	98
3.565	2.515	17.713	41.33	15.81	1.5344	8.663	2.33	2.61	0.2944	Burrow	1	4

Table 11 Physical properties of sample 3 retrieved from well H19 in the E_1f_1

Sample Length	Sample Diameter	Sample Volume	Dry Weight	Particle Volume	Pore Volume	Porosity	Particle Density	Volumetric Density	Permeability	Location	Bioturbation Index	BI Vol
(cm)	(cm)	(cm3)	(g)	(cm3)	(cm3)	(%)	(g/cm3)	(g/cm3)	(md)	(B/M)	BI	(%)
4.968	2.552	25.399	63.17	23.47	1.9337	7.6094	2.49	2.69	0.0599	Matrix	0	0
4.996	2.538	25.262	62.95	23.55	1.7086	6.7599	2.49	2.67	0.07	Burrow	5	99
4.986	2.526	24.974	63.1	23.64	1.3318	5.3299	2.53	2.67	0.0473	Burrow	3	40
4.968	2.528	24.923	62.73	23.58	1.3471	5.4024	2.52	2.66	0.0583	Burrow	1	4

Table 12 Physical properties of sample 4 retrieved from well B5 in the E_1f_2

Sample Length	Sample Diameter	Sample Volume	Dry Weight	Particle Volume	Pore Volume	Porosity	Particle Density	Volumetric Density	Permeability	Location	Bioturbation Index	BI Vol
(cm)	(cm)	(cm3)	(g)	(cm3)	(cm3)	(%)	(g/cm3)	(g/cm3)	(md)	(B/M)	BI	(%)
2.325	2.378	10.33	24.36	9.025	1.301	12.6	2.36	2.7	0.296	Burrow	5	96
2.561	2.349	11.09	26.15	9.64	1.017	9.16	2.36	2.71	0.328	Burrow	2	25
2.97	2.345	12.83	30.31	11.2	1.625	12.67	2.36	2.71	0.36	Burrow	4	90
2.39	2.371	10.55	25.02	9.321	1.03	9.76	2.37	2.7	0.246	Matrix	0	0
1.94	2.389	8.69	20.15	7.53	1.02	11.73	2.31	2.7	0.313	Matrix	0	0
2.344	2.362	10.27	24.21	9.11	1.264	12.31	2.36	2.7	0.274	Burrow	4	90
2.437	2.382	10.86	25.62	9.496	1.364	12.56	2.36	2.7	0.295	Burrow	6	100
2.911	2.353	12.66	30.11	10.652	1.353	10.69	2.38	2.83	0.301	Burrow	1	3
2.661	2.398	12.02	28.38	10.53	1.49	12.4	2.36	2.7	0.306	Matrix	0	0
2.763	2.386	12.35	29.32	9.509	2.845	23.03	2.37	3.08	0.303	Burrow	6	100
1.958	2.351	8.5	20.1	7.58	1.015	11.94	2.36	2.7	0.273	Burrow	5	94
2.331	2.382	10.39	24.41	9.018	1.37	13.19	2.35	2.71	0.243	Burrow	3	60
1.956	2.374	8.658	20.59	7.632	1.026	11.85	2.38	2.7	0.514	Burrow	6	100
2.683	2.366	11.8	27.94	10.22	1.573	13.33	2.37	2.73	0.529	Burrow	4	90
2.254	2.381	10.04	23.63	8.763	1.273	12.68	2.35	2.7	0.544	Burrow	3	50

Table 13 Physical properties of sample 5 retrieved from well B5 in the E_1f_2

Sample Length	Sample Diameter	Sample Volume	Dry Weight	Particle Volume	Pore Volume	Porosity	Particle Density	Volumetric Density	Permeability	Location	Bioturbation Index	BI Vol
(cm)	(cm)	(cm3)	(g)	(cm3)	(cm3)	(%)	(g/cm3)	(g/cm3)	(md)	(B/M)	BI	(%)
3.385	2.382	15.09	35.92	13.25	1.832	12.15	2.38	2.71	0.167	Burrow	4	90
3.205	2.377	14.22	33.94	12.5	1.771	12.45	2.39	2.71	0.133	Burrow	1	4
3.38	2.372	14.94	35.6	13.13	1.811	12.13	2.38	2.71	0.191	Burrow	2	25
2.341	2.378	10.4	24.8	9.17	1.235	11.88	2.38	2.7	0.168	Burrow	4	90
2.322	2.38	14.78	35.22	12.95	1.959	13.25	2.38	2.71	0.179	Burrow	3	40
2.302	2.381	10.25	24.42	9.01	1.201	11.72	2.38	2.71	0.139	Matrix	0	0
3.224	2.375	14.28	34.44	12.73	1.55	10.85	2.41	2.71	0.169	Burrow	6	100
2.539	2.381	11.31	26.93	10.05	1.296	11.46	2.38	2.7	0.153	Burrow	3	60
3.853	2.37	12.59	30.03	11.11	1.472	11.7	2.39	2.7	0.137	Matrix	0	0
2.76	2.385	12.33	29.55	10.95	1.379	11.18	2.4	2.7	0.152	Burrow	5	91
2.616	2.384	11.68	27.78	10.32	1.376	11.78	2.37	2.7	0.147	Burrow	2	20
2.472	2.382	11.02	26.29	9.746	1.27	11.53	2.39	2.7	0.141	Burrow	1	4

Table 14 Physical properties of sample 6 retrieved from well L5 in the E_1f_3

Sample Length	Sample Diameter	Sample Volume	Dry Weight	Particle Volume	Pore Volume	Porosity	Particle Density	Volumetric Density	Permeability	Location	Bioturbation Index	BI Vol
(cm)	(cm)	(cm3)	(g)	(cm3)	(cm3)	(%)	(g/cm3)	(g/cm3)	(md)	(B/M)	BI	(%)
0.946	2.462	4.4721	11.19	3.9276	0.5445	12.091	2.8491	2.5022	0.0111	Burrow	4	90
1.086	2.454	5.1255	12.95	4.8373	0.2883	7.511	2.6771	2.5266	0.0071	Matrix	0	0
0.972	2.452	4.6325	11.43	4.2013	0.4313	9.396	2.7206	2.4673	0.0077	Burrow	6	100
1.142	2.464	5.5048	13.64	4.8098	0.695	12.762	2.8359	2.4779	0.0089	Burrow	5	98

Table 15 Physical properties of sample 7 retrieved from well L5 in the E_1f_3

Sample Length	Sample Diameter	Sample Volume	Dry Weight	Particle Volume	Pore Volume	Porosity	Particle Density	Volumetric Density	Permeability	Location	Bioturbation Index	BI Vol
(cm)	(cm)	(cm3)	(g)	(cm3)	(cm3)	(%)	(g/cm3)	(g/cm3)	(md)	(B/M)	BI	(%)
1.722	2.478	8.3005	19.33	6.1368	2.1638	26.055	3.1499	2.3288	0.0538	Burrow	4	80
1.664	2.466	7.9434	18.89	5.7114	2.232	28.085	3.3074	2.3781	0.0477	Burrow	1	4
1.528	2.452	7.2116	17.23	5.2809	1.9307	26.759	3.2627	2.3892	0.045	Matrix	0	0
1.478	2.464	7.0441	16.72	5.0877	1.9564	27.759	3.2863	2.3736	0.0438	Burrow	2	30
1.451	2.453	6.8601	16.43	5.0473	1.9302	28.117	3.2552	2.395	0.0478	Burrow	6	100
1.455	2.454	6.8846	16.51	5.0516	1.9553	28.398	3.2683	2.3981	0.0482	Burrow	5	91

Table 16 Physical properties of sample 8 retrieved from well L5 in the E_1f_3

Sample Length	Sample Diameter	Sample Volume	Dry Weight	Particle Volume	Pore Volume	Porosity	Particle Density	Volumetric Density	Permeability	Location	Bioturbation Index	BI Vol
(cm)	(cm)	(cm3)	(g)	(cm3)	(cm3)	(%)	(g/cm3)	(g/cm3)	(md)	(B/M)	BI	(%)
3.424	2.454	16.1865	32.41	11.7102	4.4763	27.64	2.7677	2.0023	6.0245	Burrow	3	60
4.222	2.452	19.9264	40.01	14.5511	5.3753	26.962	2.7496	2.0079	7.0097	Burrow	2	30
3.523	2.451	16.6289	33.35	12.0488	4.6057	27.656	2.7679	2.0055	6.028	Burrow	5	96
3.512	2.452	16.5905	33.24	12.0112	4.5914	27.675	2.7674	2.0036	6.0321	Matrix	0	0

Table 17 Physical properties of sample 9 retrieved from well L5 in the E_1f_3

Sample Length	Sample Diameter	Sample Volume	Dry Weight	Particle Volume	Pore Volume	Porosity	Particle Density	Volumetric Density	Permeability	Location	Bioturbation Index	BI Vol
(cm)	(cm)	(cm3)	(g)	(cm3)	(cm3)	(%)	(g/cm3)	(g/cm3)	(md)	(B/M)	BI	(%)
2.552	2.462	12.143	30.06	10.7187	1.4243	11.723	2.8044	2.4755	1.1782	Burrow	6	100
3.422	2.464	16.3091	40.34	14.6921	1.6171	9.901	2.7457	2.4735	0.572	Matrix	0	0
2.816	2.442	13.194	33.17	11.8275	1.5716	11.8	2.8045	2.514	1.186	Burrow	5	98
2.515	2.461	11.968	29.62	10.5624	1.4037	11.658	2.8043	2.4749	1.172	Burrow	1	4

Acknowledgements The authors extend their gratitude to the SINOPEC Jiangsu Oilfield Company Ltd for the provision of optimal core samples and wells data. Additionally, we would like to express our appreciation to the editors and reviewers at Earth Science Informatics for their insightful feedback. Their meticulous comments and constructive remarks played a crucial role in identifying areas for improvement in our paper, enabling us to strengthen our research before publication.

Authors' contribution J.A.Q.: conceptualization, writing — original draft preparation, writing — review and editing, methodology, software, data curation, visualization, formal analysis, validation. K.S.: conceptualization, writing — original draft preparation, writing — review and editing, methodology, data curation, visualization, formal analysis. Z.J.: conceptualization, supervision, project administration, resources. C.H.: conceptualization, writing — original draft preparation, formal analysis. J.A.: writing — original draft preparation, formal analysis, software. S.A: conceptualization, writing — original draft preparation, formal analysis. B.L.: conceptualization, writing — original draft preparation, formal analysis.

Funding This study was funded by the China National Science and Technology Major Project under Grant No. 2017ZX05009-002 and the National Natural Science Foundation of China under Grant No. 41772090.

Data Availability Data sets generated during the current study are available from the corresponding author on reasonable request. The source code is made open on the GitHub repository at the link: https://github.com/JonathanAtuquayeQuaye/improved_prediction_of_Bioturbation.git\.

Code availability Improved_prediction_of_Bioturbation
Contact: jonathanatuquayequaye@gmail.com
Programme language: Python, Jupyter Notebook
Hardware required: 25,887 KB, Windows × 64, 16 GB RAM
Software required: Python 3.1.0 or higher
Programme size: 16 KB
The source code is made open on the GitHub repository at the link: https://github.com/JonathanAtuquayeQuaye/improved_prediction_of_Bioturbation.git\

Declarations

Competing Interests The authors declare no competing interests.

References

Application of neural networks to tunnel data analysis (1998) 36th AIAA Aerospace Sciences Meeting and Exhibit 26(4):373–383. https://doi.org/10.2514/6.1998-1004

Ayranci K, Yildirim IE, Waheed U bin, MacEachern JA (2021) Deep learning applications in geosciences: Insights into ichnological analysis. Appl Sci 11(16):7736

Baniak GM, Gingras MK, Pemberton SG (2013) Reservoir characterization of burrow-associated dolomites in the Upper Devonian Wabamun Group, Pine Creek gas field, Central Alberta, Canada. Mar Pet Geol 48:275–292. https://doi.org/10.1016/j.marpetgeo.2013.08.020

Beckmann M, Ebecken NFF, Pires de Lima BSL (2015) A KNN undersampling approach for data balancing. J Intell Learn Syst Appl 7(04):104–116

Birba DE (2020) A Comparative study of data splitting algorithms for machine learning model selection. Degree Project in Computer Science and Engineering 2020(1):1–23. https://www.diva-portal.org/smash/get/diva2:1506870/FULLTEXT01.pdf

Bromley RG (1996) Trace fossils: biology, taphonomy and applications. In Book (second). Chapman & Hall, London. https://doi.org/10.1017/CBO9781107415324.004

Bromley RG, Ekdale AA (1984) Chondrites: a trace fossil indicator of anoxia in sediments. Science 224(4651):872–874. https://doi.org/10.1126/science.224.4651.872

Buatois LA, Mángano MG (2007) Trace Fossils: Concepts, Problems, Prospects. Elsevier, Amsterdam

Buatois LA, Mángano MG (2011) Ichnology: Organism-substrate interactions in space and time. In: Ichnology: Organism-Substrate Interactions in Space and Time. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511975622

Chen AD (2010) Tectonic features of the Subei basin and the forming mechanism of its dustpan-shaped fault depression. Oil Gas Geol 31:140–150

Crosta AP, Souza Filho CR (1998) Spectral discrimination of mineral groups and mapping of associated alteration zones in the Carajás Mineral Province, Brazil, using remote sensing data. Econ Geol 93(8):1231–1248

Delgado R, Núñez-González JD (2019) Enhancing Confusion Entropy (CEN) for binary and multiclass classification. PLoS ONE 14(1):e0210264

Deng C (2014) The Research on Sedimentary Microfacies and Reservoir Characteristics of E1f1 Formation in Gaoji Oilfield, Jinhu Sag. China University of Petroleum (East China), Qingdao

Deshenenkov I, Polo C (2020) The digital rock analysis of biogenically induced reservoir heterogeneities in Cretaceous reservoirs of Saudi Arabia. E3S Web of Conferences 146:1–10. https://doi.org/10.1051/e3sconf/202014601004

Dorador J, Rodríguez-Tovar FJ, Hernández-Molina FJ, Stow DAV, Alvarez-Zarikian C, Acton G, Bahr A, Balestra B, Ducassou E, Flood R, Flores JA, Furota S, Grunert P, Hodell D, Jimenez-Espejo F, Kim JK, Krissek L, Kuroda J, Li B, … Xuan C (2014) Quantitative estimation of bioturbation based on digital image analysis. Mar Geol 349:55–60. https://doi.org/10.1016/j.margeo.2014.01.003

Fomel S, Liu Y (2017) Seismic image classification by deep neural networks. Geophysics 82(6):W1–W11

Gharavi A, Hassan M, Gholinezhad J, Ghoochaninejad H, Barati H, Buick J, Abbas KA (2022) International Journal of Intelligent Networks Application of machine learning techniques for identifying productive zones in unconventional reservoir. Int J Intell Netw 3(August):87–101. https://doi.org/10.1016/j.ijin.2022.08.001

Gibling MR, Bird DJ (1994) Late Paleozoic and Early Mesozoic ichnocoenoses from fluvial environments in the Sydney Basin, Nova Scotia: Palaeoecological and palaeogeographical implications. Palaeogeogr Palaeoclimatol Palaeoecol 106(1–4):105–130

Gingras MK, Pemberton SG, Saunders T (2001) The significance of bioturbation for hydrocarbon exploration. Soc Sediment Geol (Spec Publ) 69:47–64

Gingras MK, Pemberton SG, Saunders T, Clifton HE (1999) The significance of trace fossils in reservoir characterization: An example from the McMurray Formation, Alberta, Canada. AAPG Bull 83(9):1357–1375

Gu Y, Dai J (2015) Fault growth and main controlling factors in the deep area of Gaoyou Sag. Geotecton Metallog 39(1):53–61

Guinasso NL, Schink DR (1975) Quantitative estimates of biological mixing rates in abyssal sediments. J Geophys Res 80(21):3032–3043. https://doi.org/10.1029/jc080i021p03032

Hall B (2016) Facies classification using machine learning. Leading Edge 35(10):906–909. https://doi.org/10.1190/tle35100906.1

Hansen HN, Haile BG, Müller R, Jahren J (2023) New direction for regional reservoir quality prediction using machine learning - Example from the Stø Formation, SW Barents Sea, Norway. J Pet Sci Eng 220(June 2022). https://doi.org/10.1016/j.petrol.2022.111149

Hovikoski J, Gingras MK, MacEachern JA, Melnik I (2008) The impact of bioturbation on reservoir quality: A review of ichnological controls on porosity and permeability. SEPM Spec Publ 88:111–125

Huang S, Nianguang CAI, Penzuti PP, Narandes S, Wang Y, Wayne XU (2018) Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genomics Proteomics 15(1):41–51. https://doi.org/10.21873/cgp.20063

Hubert JF, Dutcher JA (2010) Scoyenia escape burrows in fluvial pebbly sand: Upper Triassic Sugarloaf Arkose, Deerfield Rift Basin, Massachusetts, USA. Ichnos 17(1):20–24. https://doi.org/10.1080/10420940903358529

Joseph VR, Vakayil A (2022) SPlit: An optimal method for data splitting. Technometrics 64(2):166–176

Li YJ, Li RF, Chen LQ, Song N, Fang J (2011) Thermal history reconstruction and hydrocarbon accumulation period discrimination of Jinhu Depression in Subei Basin. Acta Sedimentol Sin 29:395–401

Liu JH, Tang JD, Zhong SY, Yao FL, Ge ZJ (2012) Sedimentary model of lacustrine carbonate rock in the second member of Funing Formation, the West Slope of Jinhu Sag. Acta Sedimentol Sin 30(1):65–72

Liu Y, Chen Q, Hu K, Wang X, Gao F (2014) Comparison of the Bohai Bay Basin and Subei-South Yellow Sea Basin in the structural characteristics and forming mechanism. Geotecton Metallog 38(1):38–51

Liu Y, Chen Q, Wang X, Hu K, Cao S, Wu L, Gao F (2017) Influence of normal fault growth and linkage on the evolution of a rift basin: A case from the Gaoyou depression of the Subei Basin, eastern China. AAPG Bull 101(2):265–288. https://doi.org/10.1306/06281615008

Luo YH, Liu GL, Liu JH (2013) Productivity characteristics of diamictic reservoir in the second member of Funing Formation of the west slope in Jinhu Sag and its influential factors. Complex Hydrocarbon Reservoirs 6(1):38–40

Mandal PP, Rezaee R (2019) Facies classification with different machine learning algorithms – An efficient artificial intelligence technique for improved classification. Explor Geophys 2019(1):1–6. https://doi.org/10.1080/22020586.2019.12072918

Mathur A, Foody GM (2008) Multiclass and binary SVM classification: Implications for training and classification users. IEEE Geosci Remote Sens Lett 5(2):241–245

Mohammadinia F, Ranjbar A, Kafi M, Keshavarz R (2023) Application of machine learning algorithms in the classification of the flow units of the Kazhdumi reservoir in one of the oil fields in southwest Iran. J Pet Explor Prod Technol 13(6):1419–1434. https://doi.org/10.1007/s13202-023-01618-1

Mohri M, Rostamizadeh A, Talwalkar A (2012) Foundations of Machine Learning. The MIT Press Massachusettes, USA

Moldagulova A, Sulaiman RB (2017) Using the KNN algorithm for the classification of textual documents. ICIT 2017 - 8th International Conference on Information Technology, Proceedings, 665–671. https://doi.org/10.1109/ICITECH.2017.8079924

Nalepa J, Kawulok M (2019) Selecting training sets for support vector machines: a review. Artif Intell Rev 52(2):857–900

Oliver DS, Reynolds AC, Liu N (2008) Inverse theory for petroleum reservoir characterization and history matching. Inverse Theory for Petroleum Reservoir Characterization and History Matching 9780521881, 1–380. https://doi.org/10.1017/CBO9780511535642

Pemberton SG, Gingras MK (2005) Classification and characterizations of biogenically enhanced permeability. AAPG Bull 89(11):1493–1517. https://doi.org/10.1306/07050504121

Quaye JA, Jiang Z, Liu C, Adenutsi CD, Adjei S, Sarkodie K, Sokama-Neuyam YA, Lemdjou YB, Uahengo C (2023) Understanding the role of bioturbation in modifying petrophysical properties: a case from well L5 of the third-member Paleocene Funing Formation

(E1f3), Gaoyou Sag, Subei Basin, China. Arab J Geosci 16(7):1–17. https://doi.org/10.1007/s12517-023-11506-x

Quaye JA, Jiang Z, Liu C, Adenutsi CD, Boateng C D (2022) Biogenically modified reservoir rock quality: A case from the lowermost member Paleocene Funing Formation, Gaoyou Depression, Subei Basin, China. J Pet Sci Eng 111126. https://doi.org/10.1016/J.PETROL.2022.111126

Quaye JA, Jiang Z, Zhou X (2019) Bioturbation influence on reservoir rock quality: A case study of Well Bian-5 from the second member Paleocene Funing Formation in the Jinhu sag, Subei basin, China. J Petrol Sci Eng 172:1165–1173. https://doi.org/10.1016/j.petrol.2018.09.026

Raikwal JS, Saxena K (2012) Performance evaluation of SVM and k-nearest neighbor algorithm over the medical data set. Int J Comput Appl 50(14):35–39

Reineck H-E (1963) Sedimentgefüge im Bereich der südlichen Nordsee. Frankfurt am Main : Waldemar Kramer. http://lib.ugent.be/catalog/rug01:001693377

Sarma P, Gupta R (2000) Application of neural networks to petrophysical data analysis. Comput Geosci 26(4):373–383

Schäfer W (1956) Wirkungen der Benthos-Organismen auf den jungenSchichtverband. Senckenb Lethaea 37:183–263

Shao XJ, Wang CF, Huang W, Zhong SY, Liu JH, Liao GM, Chen J (2013) Sedimentary characteristics and evolutionary patterns of carbonates in Jinhu depression. Subei Basin Shiyou Xuebao/acta Petrolei Sinica 34(4):701–711. https://doi.org/10.7623/syxb201304010

Shayestegan M, Kohout J, Verešpejová L, Chovanec M, Mareš J (2024) Comparison of feature selection and supervised methods for classifying gait disorders. IEEE Access 12:17876–17894

Song N, Wang T, Chen L, Xin R (2010) Comprehensive analysis of hydrocarbon accumulation period of Upper Cretaceous Taizhou Formation in Subei Basin. Shiyou Xuebao/Acta Petrolei Sinica 31(2):180–186

Tarabulski D, Reinhardt EG (2020) Machine learning models for predicting bioturbation intensity in marine sediment cores. Mar Geol 426(106211)

Taylor AM, Goldring R (1993) Description and analysis of bioturbation and ichnofabric. J Geol Soc London 150:141–148

Tharwat A, Gaber T, Ibrahim A, Hassanien AE (2017) Linear discriminant analysis: A detailed tutorial. In AI Communications, Vol. 30(2). https://doi.org/10.3233/AIC-170729

Theng D, Bhoyar KK (2024) Feature selection techniques for machine learning: a survey of more than two decades of research. Knowl Inf Syst 66(3):1575–1637

Timmer E, Knudson C, Gingras M (2021) Applying deep learning for identifying bioturbation from core photographs. AAPG Bull 105(4):631–638

Tonkin NS, Mcllroy D, Meyer R, Moore-Turpin A (2010) Bioturbation influence on reservoir quality: A case study from the Cretaceous Ben Nevis Formation, Jeanne d'Arc Basin, offshore Newfoundland, Canada. AAPG Bull 94(7):1059–1078. https://doi.org/10.1306/12090909064

Vaibhaw SJ, Pattnaik PK (2020) Brain-computer interfaces and their applications. An Industrial IoT Approach for Pharmaceutical Industry Growth, Vol 2, 31–54. https://doi.org/10.1016/B978-0-12-821326-1.00002-4

Wang XT (2011) Structural and Sedimentary Evolution of Funing Formation in West Slop of Jinhu Sag. China University of Petroleum (East China), Qingdao, China

Worden RH, Burley SD (2003) Sandstone Diagenesis: The Evolution of Sand to Stone. In Sandstone Diagenesis (pp. 1–44). https://doi.org/10.1002/9781444304459.ch

Yang Q, Chen HY (2003) Tectonic evolution of the North Jiangsu-South Yellow Sea Basin. Pet Geol Exp 25:562–565. https://doi.org/10.3969/j.issn.1001-6112.2003.z1.008

Yi S, Yi S, Batten DJ, Yun H, Park SJ (2003) Cretaceous and Cenozoic non-marine deposits of the Northern South Yellow Sea Basin, offshore western Korea: Palynostratigraphy and palaeoenvironments. Palaeogeogr Palaeoclimatol Palaeoecol 191(1):15–44. https://doi.org/10.1016/S0031-0182(02)00637-5

Zhang T et al (2021) Application of machine learning methods for bioturbation prediction in sedimentary basins. J Geophys Eng 18(5):1021–1035

Zhang JL, Si XQ, Lin H (2006) Sedimentary Facies Analysis of the 3rd Member of the Funing Formation in the Jinhu Depression. Periodical of the Ocean University of China 36(4):535–539

Zhou X, Jiang Z, Quaye JA, Duan Y, Hu C, Liu C, Han C (2019) Ichnology and sedimentology of the trace fossil-bearing fluvial red beds from the lowermost member of the Paleocene Funing Formation in the Jinhu Depression, Subei Basin, East China. Mar Pet Geol 99:393–415. https://doi.org/10.1016/j.marpetgeo.2018.10.032

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.