

Review Paper

Artificial intelligence in petroleum and natural gas prospect evaluation and prediction

Qiao-Chu Wang^{a,b}, Dong-Xia Chen^{a,b,*}, Sha Li^{a,b}, Yu-He Chen^{a,b}, Xiang Chen^{a,b},
Mao-Sen Wang^{a,b}, Zai-Quan Yang^{a,b}, Fu-Wei Wang^{a,b}

^a State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum (Beijing), Beijing, 102249, China

^b College of Geosciences, China University of Petroleum (Beijing), Beijing, 102249, China

ARTICLE INFO

Article history:

Received 19 July 2025

Received in revised form

8 November 2025

Accepted 14 April 2026

Available online 21 April 2026

Edited by Xiu-Fang Hu

Keywords:

Artificial intelligence

Machine learning

Deep learning

Petroleum prospect

Geological facies

Petroleum exploration and exploitation

ABSTRACT

Artificial intelligence has shown great performance and considerable potential in petroleum and natural gas prospect evaluation and prediction (PPEP). Multiple machine learning (ML) and deep learning (DL) methods are applied in petroleum geology, development geology and geophysical exploration. The PPEP work in petroleum geology field which aims to provide a probability distribution for petroleum prospects with small data. The PPEP work in development field utilize seismic data as the fundamental data to predict petroleum reservoir features and favourable geological facies with the constrain of well log data. While PPEP work in geophysical exploration focus on seismic or well log facies recognition and prediction using multiple intelligent seismic denoising and well log augmentation methods. The popular AI algorithms are almost all used in all kinds of PPEP work without a unique choice and standard procedure. The decisive factors for AI application in PPEP are data quality and model interpretability. The current data size and precision used in PPEP work are hard to fulfill the requirement for AI model training, which need further augmentation and multimodal fusion. The interpretability is significant for evaluating the reliability and rationality of these data-driven models. At present, the dual-driven model driven by both of data and physical constrain may be a possible way to further decrease the requirement of data quality and increase the model interpretability. In the future, an artificial generative model which can determine petroleum prospects in multiple dimensions with a large amount of data may be the final goal for the utilization of AI in the petroleum industry.

© 2026 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Petroleum and natural gas are significant, nonrenewable fossil energy sources that are crucial for supplying energy around the world (Collett, 2002; Levorsen, 1956; Rose et al., 1984). The evaluation and prediction of petroleum prospects are important and largely determine the probability of success of petroleum exploration and exploitation (Bashkova et al., 2023; Powley, 1990).

Petroleum prospect evaluation and prediction (PPEP) work was proposed in the 1880s by White (1885). During the 1970s to the 1980s, petroleum system theory and petroleum play theory were

developed (Dow, 1974; Lee and Wang, 1983; Perrodon and Masse, 1984). Based on these theories, multiple kinds of PPEP methods have been introduced and applied through quantitative evaluation of the key factors of petroleum formation processes (Cox, 1993; Magoon and Dow, 1992; Reed et al., 1989; Singer, 1993). From the 1990s to the 2000s, many giant oil and gas fields were discovered, which accelerated the process of industrialization and modernization around the world (Agterberg, 2011; Van Der Meer et al., 2002). In the 21st century, unconventional petroleum resources have shown high resource potential and exploration value (Gao et al., 2024; Miletić et al., 2024; Zangqa et al., 2024). The large difference between conventional and unconventional oil and gas resources has led to the poor performance of traditional PPEP methods in unconventional resource exploration (Allan, 1989; Voordouw, 2011; Wang et al., 2023; Zhen et al., 2022). The large

* Corresponding author.

E-mail address: lindachen@cup.edu.cn (D.-X. Chen).

Peer review under the responsibility of China University of Petroleum (Beijing).

disparity in well production presents a major challenge for PPEP work.

In the course of the evolution of information technology, the concept of intelligence and artificial intelligence (AI) technology have been applied to all fields and are changing the world (Dorotic et al., 2024; Gafni et al., 2024; Pradhan, 2013; Wang et al., 2017). At present, the data-based intelligent era is beginning (Fig. 1).

This strong data analysis and calculation capability and self-improving mechanism highlight the importance of AI technology in scientific research and industrialized applications, including PPEP work (Bengio et al., 1994; Hastie et al., 2008; Imamverdiyev and Sukhostat, 2019). First, the evaluation and prediction models constructed by AI algorithms are based entirely on the original dataset, and this objective process may reveal the actual relationship between the data and excavate the principles (Cortes and Vapnik, 1995; Hinton et al., 2011; Quinlan, 2003). Second, the theories applied to multiple AI methods are available to solve complex nonlinear problems that cannot always be solved with statistical methods (Denton et al., 2015; Le Roux and Bengio, 2008; Vapnik et al., 1996). Third, the automatic modelling process and self-improving mechanism make it possible to efficiently determine the best model among a large number of constructed models (usually thousands or millions) in a short time, which is impossible for traditional PPEP methods (Hinton et al., 2006; Mnih et al., 2015; Szegedy et al., 2015). However, the application of AI in PPEP faces many challenges, which concern the size and quality of the database, the conflicts between petroleum geological knowledge and the results of pure AI models, and the predicting capability and generalization of AI under the background of limited data (Han et al., 2020; Wang et al., 2023).

The remainder of this paper is organized as follows. First, the concepts and techniques of AI applied in PPEP are introduced briefly. Second, the PPEP methods, which are applied in petroleum geology, development geology and geophysical exploration fields, are reviewed for their algorithm selection, model performance, model limitation and optimization. Third, the accuracy, applicability, generalizability and practicality of AI-based PPEP methods

are discussed. Finally, we offer a perspective on the future development of AI in PPEP work and even in the petroleum industry.

2. Artificial intelligence concepts and techniques

In this section, the dominant AI concepts and techniques related to PPEP research and industry are reviewed, including the principles of the algorithms, the advantages and disadvantages of the algorithms for PPEP and their adaptability. In accordance with the principles and development history of AI, intelligent methods are introduced in the order of machine learning (ML), ensemble learning (EL) and deep learning (DL). Note that deep learning and ensemble learning are also machine learning methods, although these different concepts represent three different development stages of AI for application in the petroleum exploration and exploitation industry. Therefore, we distinguish EL and DL from ML and introduce them separately.

2.1. General machine learning

2.1.1. Decision tree models

The decision tree algorithm is one of the basic algorithms in ML. A decision tree is a hierarchical model that determines decision rules to separate the input variables into different homogeneous subgroups (Quinlan, 1986). There are several algorithms for decision tree (DT) construction, including ID3, C4.5 and classification and regression tree (CART), which are more popular in PPEP work because they are binary trees with two categories for the target variable (whether the petroleum prospect is favourable or not) (Breiman et al., 1984; Quinlan, 1986). The DT model is easily to be overfitting in PPEP application, and there are several ways to perform overfitting calibration, including cost-complexity pruning, minimum samples per leaf, and minimizing the Gini index (Chan et al., 2021; Huo et al., 2021).

The DT algorithm is one of the most common methods with high efficiency and high interpretability, which can be selected as

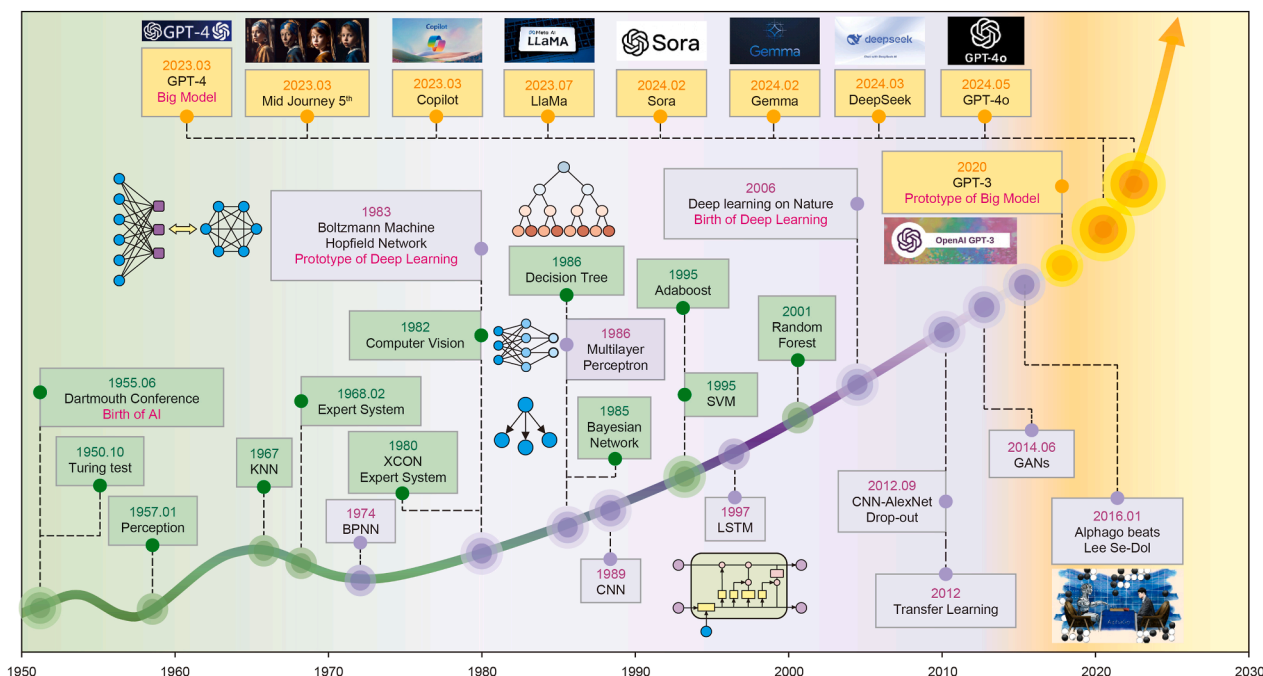


Fig. 1. The history of artificial intelligence development.

the initial model for almost any field. However, the DT model is sensitive to noise in training data, which will lead to overfitting and weak robustness.

2.1.2. Support vector machine models

The support vector algorithm is a nonlinear generalization algorithm that is powerful for optical character recognition (Boser et al., 1992; Guyon et al., 1992; Vapnik, 2006; Vapnik et al., 1996). The support vector machine (SVM) can address both classification and regression problems via support vector classification (SVC) and support vector regression (SVR) (Burges, 1998; Drucker et al., 1997). The key to applying the SVM is to find a hyperplane in a high- or infinite-dimensional space, which plays different roles in SVC and SVR (Chang and Lin, 2001; Cortes and Vapnik, 1995).

For an SVC model, the best classifier depends on the largest distance to the nearest training data point of any class (Fig. 2(a)) (Huber, 1981; Keerthi et al., 2001; Lee and Mangasarian, 2001). For an SVR model, the hyperplane being searched is the one with the largest deviation ε from the farthest data point in the training set (Fig. 2(b)) (Bennett and Mangasarian, 1992; Cortes and Vapnik, 1995). The kernel functions are also important for the SVM, which can significantly enhance the performance of SVM models. Common kernel functions include linear functions, polynomial functions, radial basis functions (RBFs), and sigmoid functions (Boser et al., 1992; Burges, 1998; Cortes and Vapnik, 1995; Platt, 1998). Compared to DT, the application of hyperplane and kernel functions can substantially enhance the model robustness, especially in classification problems for small-sample with high-dimensional data.

2.1.3. K-means models

K-means analysis is a typical cluster analysis method that is an unsupervised learning algorithm (Ankerst et al., 1999; MacQueen, 1967; Martin et al., 1996). The K-means algorithm is applied to separate the input dataset into several groups named clusters with genetic relationships but without labels via the determination of centroids (Charles, 2003; Cheng, 1995). It is one of the most popular cluster analysis methods and is also used for dimensionality reduction and vector quantization (Charles, 2003; Dempster et al., 1977; Dunn, 1973). The model generates k centroids via random sampling (Fig. 3(a)); then each data point is assigned to the centroid at the shorter distance via the Euclidean distance formula (Fig. 3(b)). The new centroid of a certain cluster is defined and

replaces the original centroid by calculating the mean value of all the data points belonging to that cluster (Fig. 3(c)). The iteration process ends when the location of the centroid no longer changes (Fig. 3(d)) (Charles, 2003).

The biggest advantage of the K-means model is its universality as an unsupervised algorithm, which does not need labels for model training. Such unsupervised method is of great significance for PPEP with limited drilling or seismic data. However, without supervision and guidance, the model may provide the results which do not totally obey the petroleum geological knowledge (Comaniciu and Meer, 2002).

2.1.4. Naïve Bayes models

Naïve Bayes (NB) theory is based on probability theory (Jensen, 1996). Generally, the probability calculated by naïve Bayes methods fits the naïve Bayes assumption: the feature variables are conditionally independent of each other. According to this assumption, binary classification problems can be solved via the following steps with naïve Bayes theory (Friedman et al., 1997):

- (1) For the target variable Y and feature variable X , if the probability of $Y = 1$ is required, then:

$$P(Y = 1|X) = \frac{P(X|Y = 1) \times P(Y = 1)}{P(X)} \quad (1)$$

- (2) If we have i feature variables, then:

$$X = \{X_1 = x_1, X_2 = x_2 \dots X_n = x_n\} \quad (2)$$

The value of $P(Y = 1|X)$ can be calculated as follows:

$$P(X|Y = 1) = \prod_{i=1}^n P(X_i = x_i|Y = 1) \quad (3)$$

The value of $P(X)$ for $P(Y = 1)$ can be calculated as follows:

$$P(X) = P(Y = 1) \times P(X|Y = 1) + P(Y = 0) \times P(X|Y = 0) \quad (4)$$

Then, the value of $P(Y = 1|X)$ can be determined via a combination of Eqs. (1) and (2). The NB algorithm is one of the basic ML methods based on a statistical principle, which also has high interpretability. However, owing to its assumption, this algorithm can only provide good performance for situations that all the feature variables are independent without relation, which is almost impossible for PPEP. In other words, the application of NB

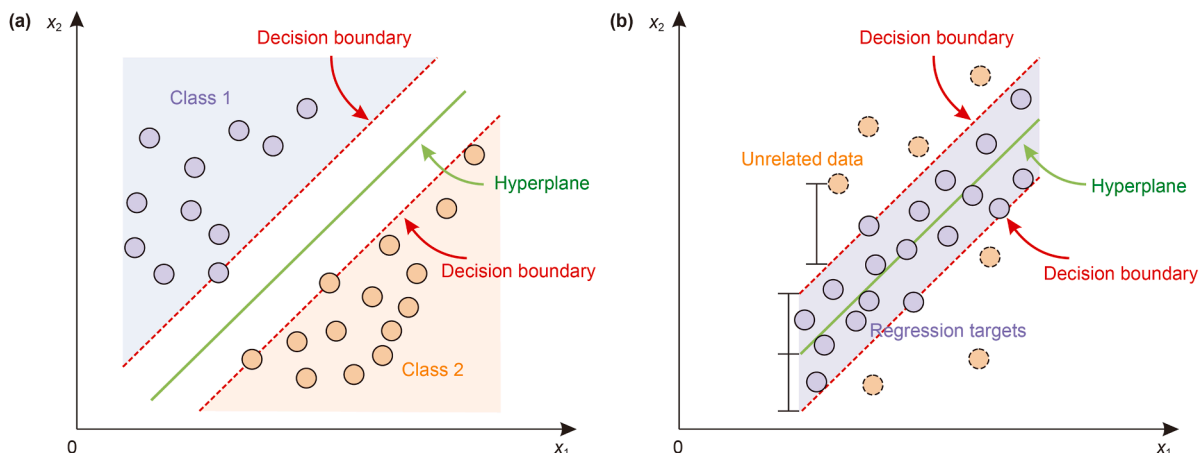


Fig. 2. The workflow of the support vector machine (Cortes and Vapnik, 1995). (a) Support vector classification; (b) support vector regression.

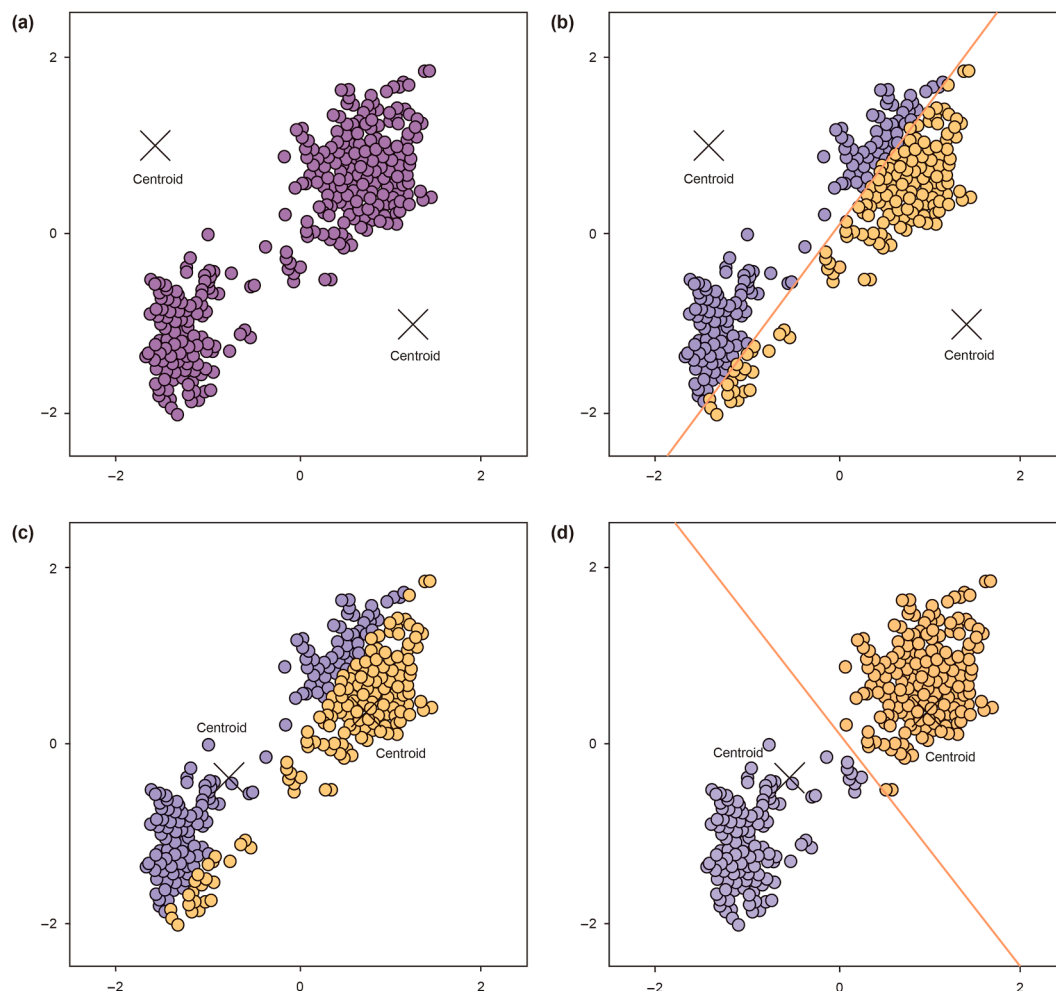


Fig. 3. The workflow of a K-means cluster.

in PPEP requires the feature selection processes to select the independent geological features. Otherwise, the features may need to be transferred to independent ones by PCA or other ways.

2.2. Deep learning

2.2.1. Backpropagation neural network (BPNN) models

The BPNN is one of the most popular DL algorithms and demonstrates good performance. The BPNN is characterized by a self-optimizing ability through error-back propagation, which allows the mistake to be transferred back to the network and guides the model to change the parameters and hence improve the model performance (Fig. 4). Compared to the general ML algorithms, the BPNN model in PPEP can easily obtain higher accuracy and better performance with sufficient data. The self-optimizing procedure also leads to higher efficiency of model optimization. While as one of the most important DL algorithms, the BPNN is sensitive to noisy data, which means it may be difficult to be directly applied to geophysical field because of the large amount of noisy in seismic data. The fully-connected structure will also lead to local optima instead of good generalization, which will also be hard to be

applied in the region with less drilling wells (Hornik et al., 1989, 1990).

2.2.2. Convolutional neural network (CNN) models

The CNN is one of the most important neural network DL algorithms for image segmentation and classification (Alex et al., 2012; LeCun et al., 1989, 1990, 1998). Existing CNNs have different architectures for dealing with different types of input data. Generally, CNNs focus on 2D image segmentation and classification; however, at present, CNNs for 3D data are also widely applied in many fields. The key to CNN is converting images into digital information, which is like a quantification process of image data that can extract sufficient information hiding in the details of the image and hence provide the evidence for PPEP in favourable zone identification, effective reservoir determination and pore structure reconstruction.

Comparing to other artificial neural networks, the convolutional kernels have translation invariance, and pooling layers can also enhance noise resistance by feature extraction, which means CNN is an excellent solution for PPEP using geophysical technologies. The CNN is sensitive to the image transformation including rotation and scaling. The parameter sharing mechanism can also

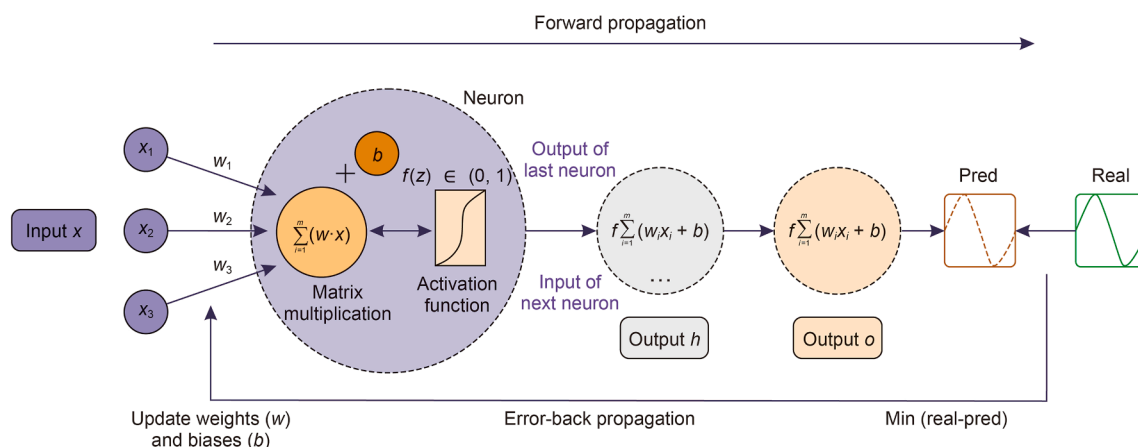


Fig. 4. The workflow of a BPNN.

reduce overfitting risk and lead to good robustness for CNN model. Furthermore, data augmentation and transfer learning strategy can also further optimize the CNN model.

2.2.3. Generative adversarial network (GAN) models

A GAN model consists of two independent neural networks. The generative model (G) is applied for capturing data and generating images, and the role of the discriminator (D) is to estimate the probability that a given distribution or image is from the real training examples rather than from G (Goodfellow et al., 2014; Mirza and Osindero, 2014). Similar to CNN, GANs is also good at PPEP based on images. While the GANs are characterized by its flexibility because they can be a supervised, semi-supervised and even an unsupervised model. In this case, GANs can be applied in many scenarios with different data conditions and PPEP tasks. Furthermore, the interpretability of GANs and CNN is better than that of other ML or DL methods because they can show the images through pooling or convolution processes. Compared to CNN, the GANs may provide more creative results because of its unsupervised property.

Owing to the adversarial module of GANs, the whole framework is unstable and has a risk of collapse during the training process, which is the largest problem in GAN applications. The generalization of GANs largely depends on the diversity of the data (Denton et al., 2015). There are several ways to enhance the robustness of GANs, including progressive training, two-timescale update, and diversity loss function.

2.2.4. Recurrent neural network (RNN) models

Recurrent Neural Networks (RNNs) are defined as a class of artificial neural networks specifically engineered for processing sequential data. Distinguished from feedforward architectures,

RNNs incorporate cyclic connections that enable dynamic temporal behaviour and state persistence, making them particularly suitable for time-series analysis, natural language processing, and speech recognition (Fig. 5). For any scientific field related to sequential problems, RNNs can be applied for predicting and evaluating work. Considering that well log and seismic data are both sequential data, the RNN can be used for PPEP in many concrete problems, especially in the scenarios of predicting while drilling (Choromanska et al., 2014). While constrained by the structure, the RNNs only has a short memory, which means it can only make decisions based on data adjacent to the target. LSTM and GRU have solved this problem to some extent, but they are still characterized by local optimum. In many cases of PPEP, the global optimum is our final requirement because of the complex geological feature and intensive heterogeneity of the petroleum reservoir require a global sight.

2.2.5. Transformer models

Transformer framework is one of the most popular and effective frameworks in DL in recent years, which exactly overcomes the disadvantage of traditional RNNs in local optima with limited information. It is based on self-attention mechanism which has promoted the development of large language models (LLM). The transformer framework is now widely used in translation, text generation, image detection, etc.

The typical transformer model consists of an encoder-decoder structure (Fig. 6). The encoder maps an input sequence of symbol representations to a continuous sequence. For the sequence given by the encoder, the decoder will generate the corresponding output sequence. The whole process is auto-regressive. During this process, the attention function is applied in the Multi-Head Attention module. This attention function is described as

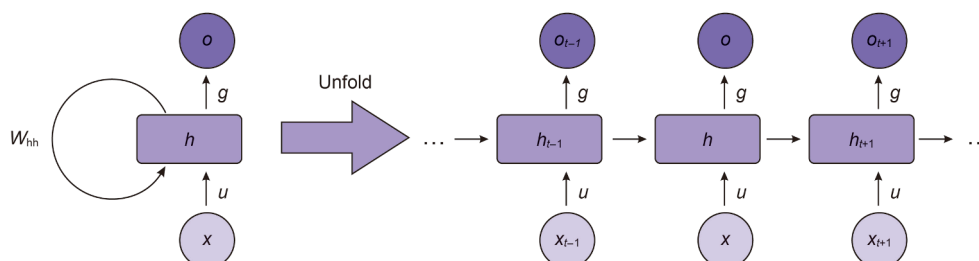


Fig. 5. The architecture of RNNs.

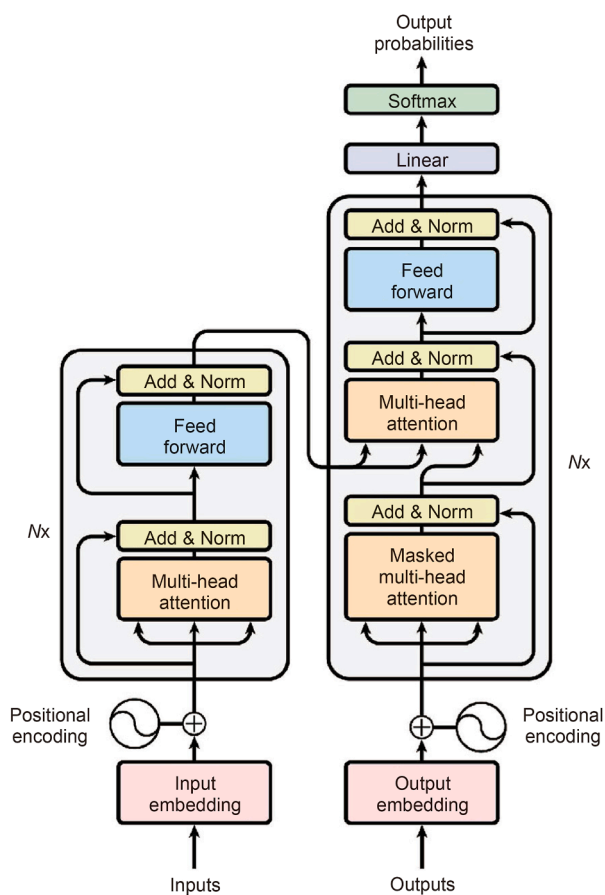


Fig. 6. The architecture of transformer framework.

mapping a query and a set of key-value pairs to an output, among which the query (**Q**), key (**K**), value (**V**) and output are all vectors. The final output is computed as a weight of values, and the value is computed by compatibility of the query with the corresponding key:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (5)$$

As a global modelling algorithm with self-attention mechanism, the transformer model has strong ability to resist local interference and has excellent robustness compared to other RNN architectures. In theory, transformer architecture can be applied to any field of petroleum geology with ideal results and generalization ability. However, considering the data size and high-computation equipment that the transformer model requires, the application of transformer model undoubtedly has a high threshold in PPEP. Generally, a pre-training process on large amount of data will generate a powerful universal model, which is then optimized by fine-tuning and applied to concrete scenarios.

2.3. Ensemble learning

Ensemble learning combines several weak learners (ML models) with poor performance into a strong learner with good performance (Breiman, 2001; Kam, 1995). Therefore, ensemble learning serves mostly to enhance the performance of existing ML models instead of creating new algorithms (Freund, 1995, 1999; Friedman, 2001). Adaptive boosting (AdaBoost), extreme gradient boosting (XgBoost), gradient boosting decision tree

(GBDT) and random forest (RF) are typical ensemble learning models for enhancing the performance of DT models (Breiman, 2001; Chen et al., 2016; Freund, 1995; Friedman et al., 2000; Kamiński et al., 2018). Note that the ensemble learning models mentioned in this manuscript are the conventional EL models based on ML without the ensemble DL methods.

2.3.1. Bagging models

The bagging model is actually a voting system. For a given problem based on the same dataset, a series of different general machine learning models can be constructed by different compositions of training, validation, and testing data determined by random sampling (Breiman, 2001; Kim, 2016). The final output of the classification or regression problem is determined by the distribution of all the outputs of each single model (Fig. 7).

2.3.2. Boosting models

The boosting model is another way to enhance the performance of a weak single ML model (Freund, 1995, 1999). When the first model is constructed, the data points incorrectly predicted by such a model are given more attention by increasing the weight of these data and reconstructing the model based on the processed dataset. In this way, the model can be fixed and improved compared to the former model. After a certain number of iterations, the final model will show satisfactory performance (Fig. 8) (Chen et al., 2016; Kim, 2016).

3. Artificial intelligence in PPEP work

In the 2000s, ML methods were first applied in PPEP procedures, which were based mainly on NB (Chen and Osadetz, 2006; Martinelli et al., 2011; Van Wees et al., 2008). DT and SVM were subsequently applied for PPEP in petroliferous basins (Armitage and Stright, 2010; Virieux and Operto, 2009; Wang et al., 2023). In the early 21st century, neural networks were widely used in the petroleum industry for petroleum reservoir modelling and then for exploration target determination. In petroleum industrial system, the PPEP is achieved by petroleum geology, development geology and geophysical exploration with different scopes, data sizes and types, and concrete predicting and evaluating targets.

3.1. AI application in petroleum geology

The PPEP by petroleum geological methods are characterized by relatively larger scope with high data dimension and small data size. The target of petroleum geological evaluation is a petroleum reservoir trap, which means we will only obtain dozens to hundreds of data points. In this case, the data size is always the biggest problem in AI application.

3.1.1. Gas layer probability prediction by NB

The NB algorithm was used for PPEP in the early stage of petroleum exploration and exploitation. With the provision of parameters related to hydrocarbon accumulation and oil and gas reservoir formation processes and the discovery of petroleum prospect messages, an original dataset with labels can be obtained (Martinelli et al., 2011; Jensen, 1996). Specifically, the labels can be obtained from the productivity and well interpretation results of drilled wells or the hydrocarbon traps discovered. Owing to differences in labels, the result of the NB model can also be the potential petroleum prospects or the locations of drilling wells (Chen and Osadetz, 2006). The input data include the node attribute, which contains the source, prospect, and segment (Chen and Osadetz, 2006; Van Wees et al., 2008), and any quantitative

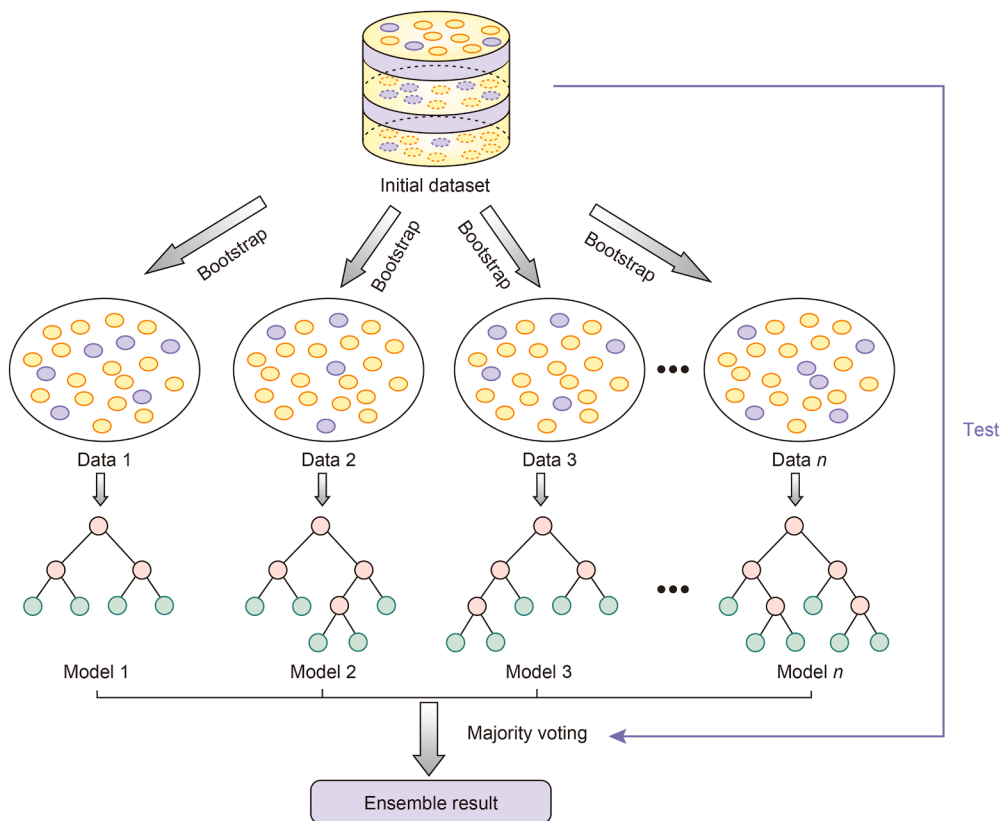


Fig. 7. The workflow of a bagging model (Kim, 2016).

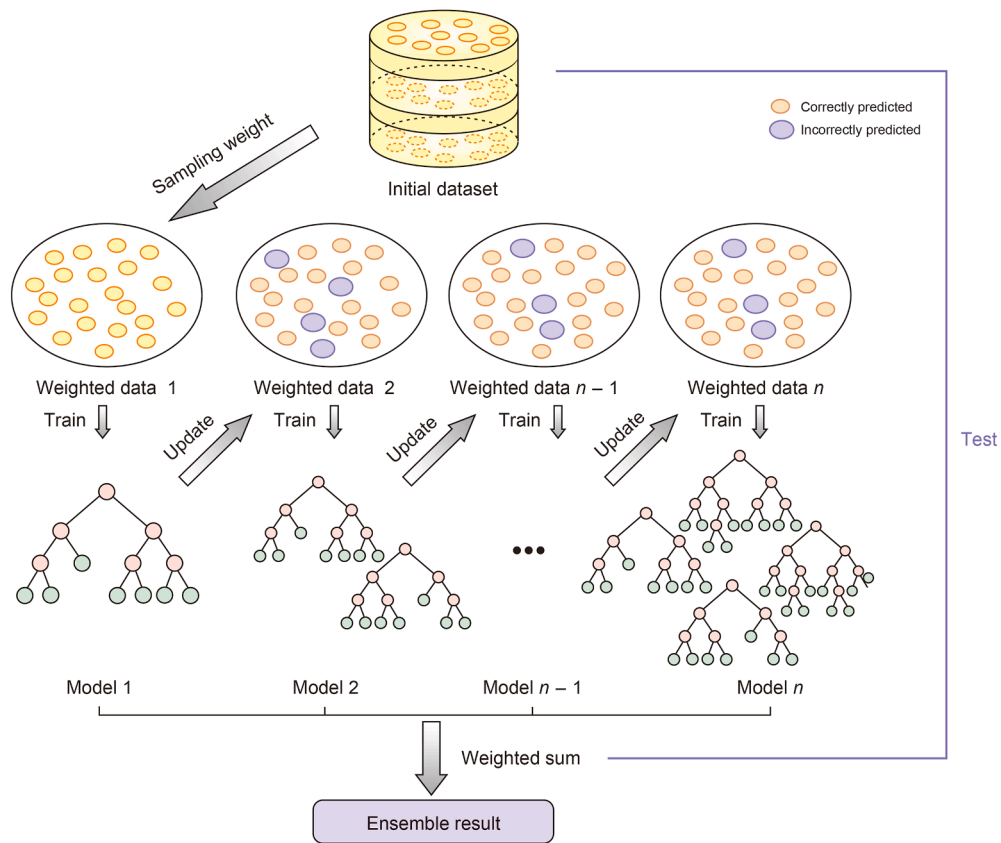


Fig. 8. The workflow of a boosting model (Kim, 2016).

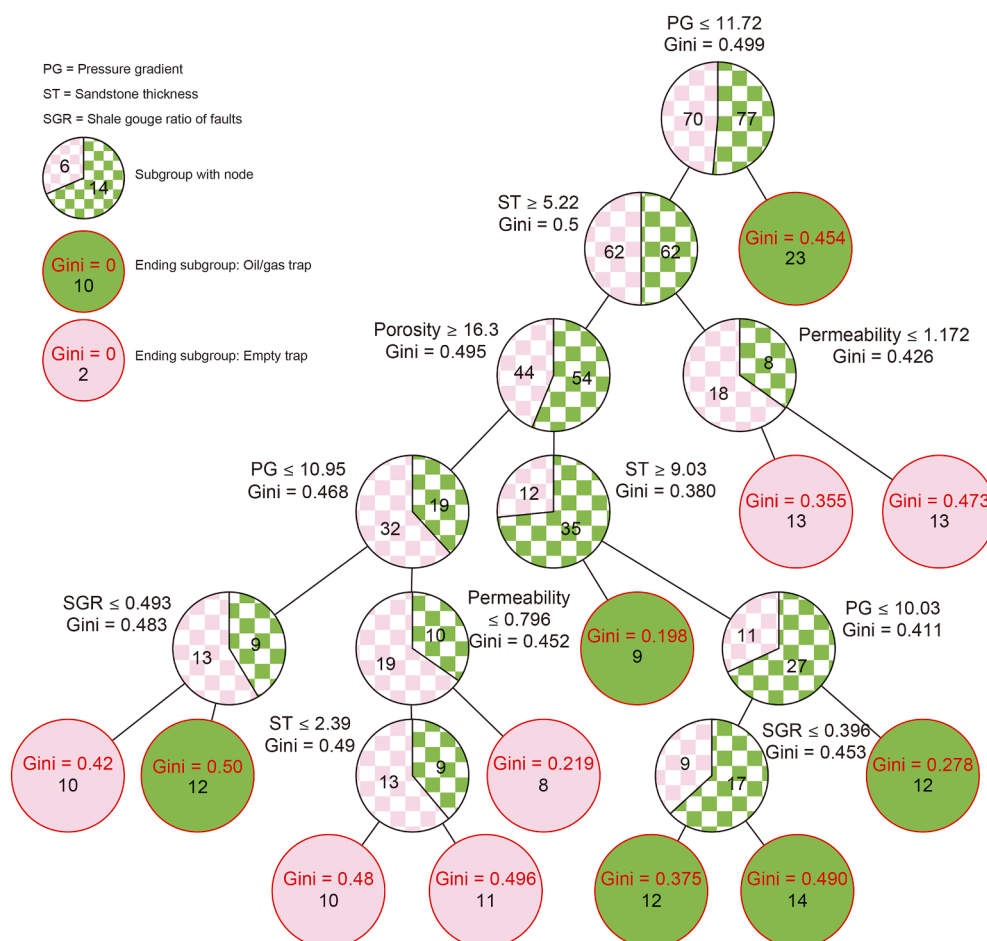


Fig. 11. DT model for PPEP work (Wang et al., 2023).

first problem, the calibration process named pruning is necessary for all DT models (Akkaş et al., 2015; Ma et al., 2022). The dominant hyperparameters for pruning include the maximum depth, minimum samples for splitting, and minimum samples per subgroup (Lawson et al., 2017; Sun et al., 2022). Considering the limited data, maximum depth is the most effective hyperparameter to keep as many samples as possible. For the second problem, cross-validation is a general way to check the performance of the DT model with small data (Gjelsvik et al., 2023; Kelter, 2021; Wainer and Cawley, 2021). The synthetic minority over-sampling technique (SMOTE) is also a representative over-sampling method for class imbalance of the data, which often occurs in PPEP related dataset. The SMOTE generates data by combining the selected data and its randomly adjacent data points, which can overcome the overfitting and keep the integrity of the original data. Generally, the oversampling methods can significantly improve the performance of the model in total accuracy and sensitivity. Furthermore, some promoted DT structures, such as the RUSBoosted tree, can also be helpful for small and unbalanced datasets (Ma et al., 2022).

3.1.3. Favourable petroleum prospect determination by SVC

As an advanced algorithm, the SVM has been shown to be applicable for PPEP at both large and small scales (Bennett and Mangasarian, 1992; Cortes and Vapnik, 1995; Keerthi et al., 2001; Wahba, 1998). The SVC model construction process for PPEP in petroleum geology is similar to that of the DT. However, the priority of the SVC is classification, which makes binary classification easy because of the increasing dimensionality and hyperplane

theory. By applying the SVC, an effective PPEP model with good performance and strong generalization ability can be constructed with a relatively small dataset (Wang et al., 2023). Wang et al. (2023) applied the SVC model to evaluate and predict favourable zones in the Dongpu depression using 187 pairs of hydrocarbon trap data. The final result indicated a prediction total accuracy of 88.92%, precision of 0.88, and F1-score of 0.91, which is high enough for basin-scale petroleum prospect exploration (Fig. 12).

3.1.4. Favourable petroleum prospect determination by EL

Constrained by the limited data, the single AI model applied in petroleum geology cannot always perform well, while EL is an ideal way to enhance the model performance and robustness. Furthermore, the relatively small data size also indicates that even if we need to train multiple models, this process won't require too much time or too expensive hardware.

In this case, the ensemble PPEP model is consistently applied for large-scale PPEP. Owing to the relatively small dataset, the utilization of the bagging and boosting models can significantly enhance the performance of a single model in less than 1 GPU hour (Fang et al., 2024; Saikia et al., 2020) (Fig. 13). The final result showed a predicting accuracy improvement of approximate 23% using the boosted model for predicting the petroleum prospect, which is an effective strategy for AI applications in petroleum geology.

To conclude, the AI application to PPEP by petroleum geological analysis concentrates on general ML methods including NB, DT and SVC owing to the relatively small data size of petroleum geology

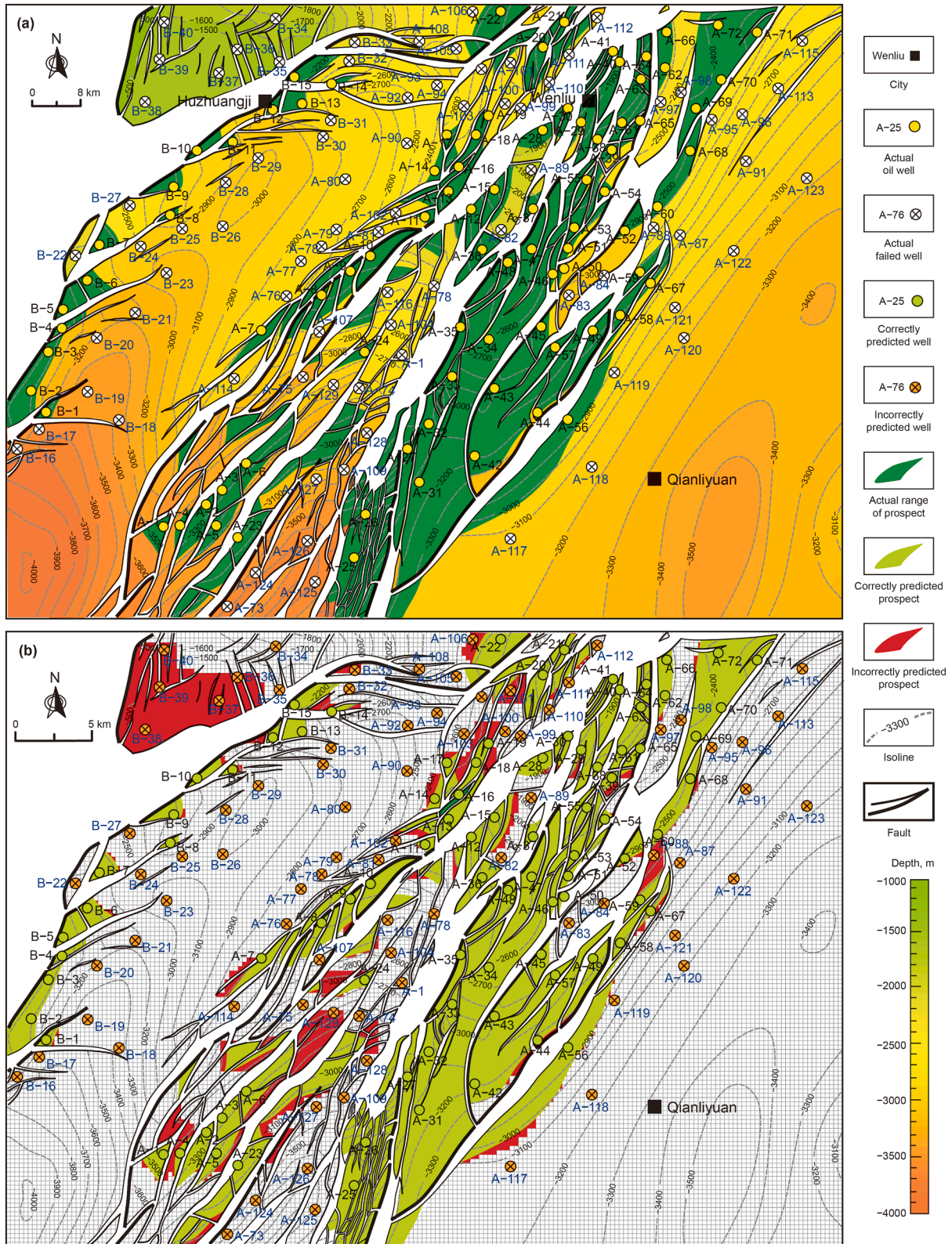


Fig. 12. Favourable zone determination via the SVC (Wang et al., 2023). (a) The actual distributions of petroleum prospects (oil reservoirs) and oil/failed drilling wells; (b) the prediction results of the petroleum prospects via the SVC with the correctly and incorrectly predicted drilling well locations.

data. DT has the best visualization ability, while its overfitting and weak robustness make it difficult to be applied for prediction work. SVC is suitable for PPEP work in the aspect of petroleum geology with the hyperplane theory. EL is an effective way to

enhance the accuracy and robustness of general ML models without too much time cost. NB is no longer suitable for PPEP work owing to the independent feature assumption (Table 1). Furthermore, the feature importance can be obtained by DT and SVM,

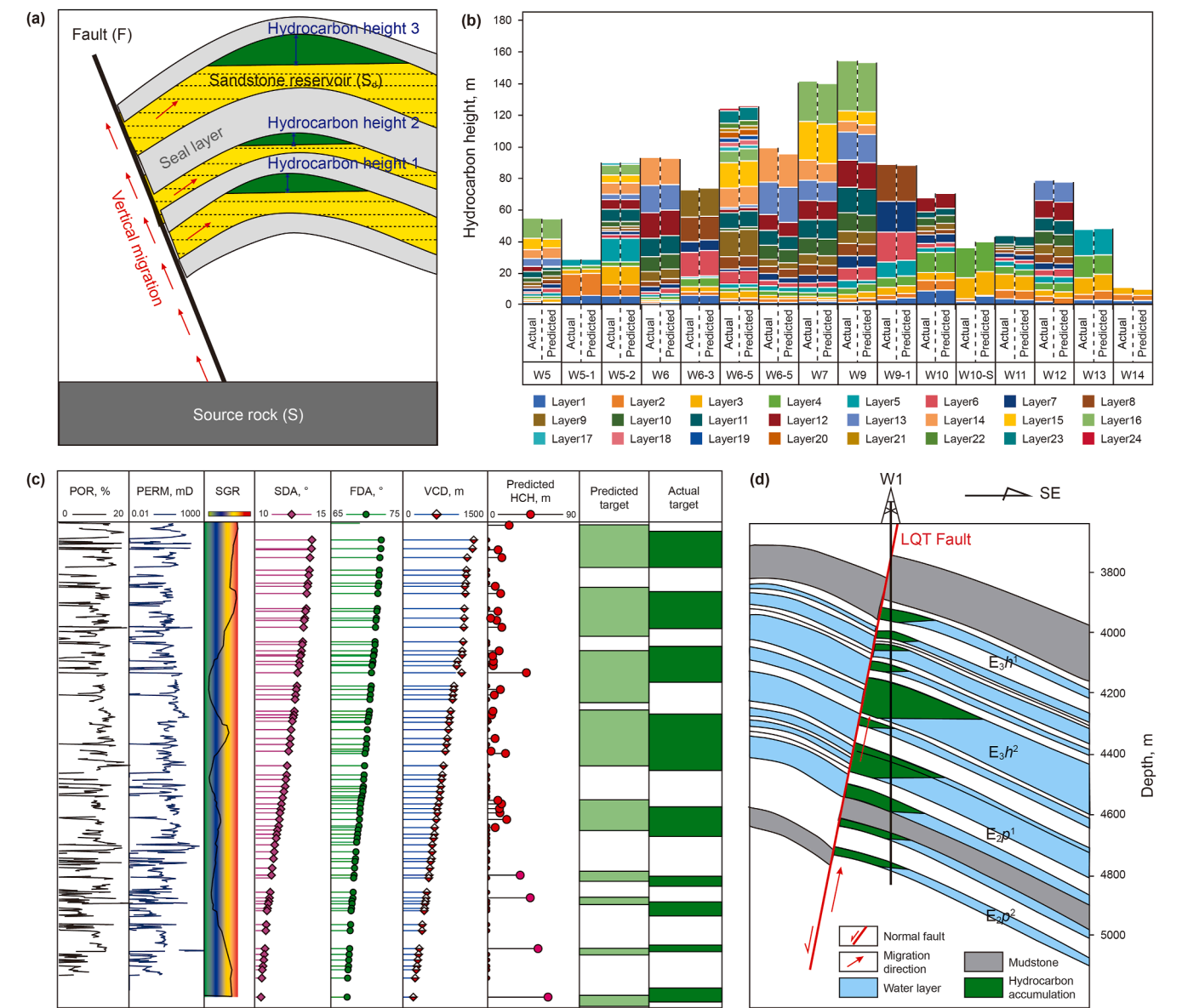


Fig. 13. Hydrocarbon height prediction via ensemble neural networks (Wang et al., 2024). (a) The definition of hydrocarbon height in petroleum exploration; (b) the actual and predicted hydrocarbon heights of the oil reservoirs; (c) the application of ensemble learning models for petroleum prospect prediction for the W1 well; (d) the actual petroleum prospect distributions for the W1 well block.

Table 1
Comparison of ML models applied for PPEP in petroleum geology.

Algorithm	Target	Data type	Data size	Accuracy	Advantage	Limitation
NB	Gas layer probability	Well log	<100	70%–80%	Low cost	Accuracy; unsuitable assumption
DT	Hydrocarbon accumulation probability	Laboratory test, well log, seismic interpretation	150–200	65%–88%	Visualization; high efficiency	Overfitting; weak generalization
SVC	Hydrocarbon accumulation probability	Laboratory test, well log, seismic interpretation	500–1000	84%–92%	High accuracy; high efficiency	Visualization; robustness
EL	Hydrocarbon accumulation probability	Well log	5000–10000	92%–95%	High accuracy; good robustness	Visualization; high cost

which can be regarded as the controlling factor analysis of hydrocarbon accumulation. Such quantitative analysis is also an ideal way to verify the rationality of the ML results by professional petroleum geological knowledge.

In some situations, the total accuracy cannot reflect the actual performance of a ML model, for which different evaluation metrics are introduced (Quinlan, 2003). In the context of PPEP, the evaluation criteria may have more

practical significance for petroleum exploration strategies and decision-making (Table 1). The focus on different criteria reflects the different requirements for petroleum exploration.

3.2. AI application in development geology

AI application in development geology is different from that in petroleum geology in research scale, data origin and evaluating targets (Adnan et al., 2024; Saikia et al., 2020; Zhao et al., 2023). Based on the petroleum geological theory, the target of PPEP by development geology focuses on the favourable petroleum reservoirs with the thickness of several meters. With the smaller research scale, well log and seismic data are the dominant data which size is larger than that applied in petroleum geology (Avalos and Ortiz, 2019; Yue et al., 2019; Zhang et al., 2020). In development geology works, the well log data is always regarded as the supervised label or the validation of the model, while the predicting work primarily relies on seismic data. Along with the exploration of unconventional reservoirs, which are characterized by intensive heterogeneity, the PPEP model in development geology requires more drilling wells for validation test. With limited well log data, the efficiency of the model cannot be guaranteed. In other words, how to apply the data-driven ML and DL model in an unconventional resource exploration field (characterized by insufficient well log data) is not only the main problem in AI application in PPEP, but also the primary problem in petroleum industry.

3.2.1. Hydrocarbon reservoir prediction by SVR

In the development geology field, 2D–3D seismic datasets can provide enough effective data for an SVR to determine the oil and gas prospects by predicting physical properties of sand bodies including sandstone thickness, porosity, permeability, etc. (Fig. 14(a)) (Al-Anazi and Gates, 2010; Demyanov et al., 2015). In addition, the well-log curves and the corresponding parameters extracted from the well-log data are applied as the supervised data to constrain the prediction results (Fig. 14(b)) (Gjelsvik et al., 2023; Li et al., 2020; Yue et al., 2019). However, regression model training is more difficult than classification model training. In this case, an optimizing procedure is required. For example, Li et al. (2020) applied adjacent reservoir thickness to further constrain the predicted reservoir thickness using development geological statistic conclusions. The final correlation coefficient of the model reached 0.806, which significantly enhanced the predicting accuracy of the sandstone thickness (Fig. 14(c)).

3.2.2. Favourable reservoir model construction by GAN

The SVC can provide satisfactory results using the supervised well log data. However, in new exploration areas or deep strata and deep-sea areas, which are characterized by fewer well log data, it is difficult to gather supervised data to construct an effective PPEP model (Song et al., 2021, 2022). As an unsupervised algorithm, the GAN is introduced in PPEP for reservoir modelling, geological facies modelling, and for the favourable zone determination in development geological analysis (Sun et al., 2023).

For PPEP by the GAN, the generator is applied to generate reservoir models or geological facies models that match real models underground, and the role of the discriminator is to discriminate the models generated by the generator with the help of real facies and reservoir models in 2D and 3D (Fig. 15(a)) (Linde et al., 2015). Generally, the input data for the discriminator can include images provided by satellites, outcrop pictures, dominant typical facies maps or reservoir distribution maps in 2D and 3D with different attributes (Fig. 15(b) and (c)) (Goodfellow et al.,

2014; Jiang et al., 2023; Song et al., 2021). With a high-performance GPU, models for PPEP can be constructed in different patterns when the loss of the generator and the discriminator reach a balance (Fig. 15(d)). This procedure will cost several to hundreds of GPU hours according to the data size. The key to the practicability of the GAN model is the calibration process. The most common method for testing and calibrating the GAN model are the “post-GAN” approach and Markov chain Monte Carlo (MCMC) algorithms. Such methods will record the loss between generated images and real ones and construct a loss function. Then the latent vector will be set with the aim of decreasing such loss until it is small enough (Laloy et al., 2018; Zhang et al., 2019).

Besides the data size, the geological rationality and training stability are main problems owing to the random generating procedure. Without sufficient well log data as supervised labels, the model with obvious mistake that violated the common rules or the training instability and mode collapse will largely decrease the practicality of GANs. In fact, well log data are not the only restricting conditions and supervised labels for geological facies or reservoirs. As real or formerly existing geological bodies, facies and reservoir models should follow some natural rules, namely, the global features of rivers, mountains, lakes, bars, channels, caves, etc. (Song et al., 2022).

Considering this point, Song et al. (2021) introduced an improved GAN model, GANSim. The improvement of GANSim is to put the well interpretation data, channel probability map, and global features of channels (characterized by the parameter of the channel complex proportion) into the generator as restricted conditions that the generated model must follow (Fig. 16(a)) (Song et al., 2021). Such an operation will achieve the constraining of generating image by geological knowledge and global feature of geological bodies (Fig. 16(b)). Compared with the traditional GANs model, GANSim has high efficiency with a lower time cost and high accuracy with the correlation coefficient of 0.99 in channel facies prediction (Fig. 16(c)). Furthermore, the application of multiple supervised information significantly shortened the training time from several hundreds of GPU hours to less than 100 GPU hours without frequent mode collapsing.

To summarize, the SVR is effective for PPEP by development geological analysis with high-quality seismic data and sufficient well log data as the supervised label. While with less drilling wells, the GANs model can be applied for PPEP due to its ability of generating, which is a dominate method for study area with insufficient fundamental data. However, the randomly generating process not only leads to meaningless outputs but also tend to cause mode collapsing. In this case, GANSim with multiple restricting conditions which do not rely on well log but on global features is an ideal solution for the disadvantages of normal GANs (Table 2).

3.3. AI application in geophysical exploration

The data origin of AI application is also seismic and well log data, which is similar to that of development geology. While the PPEP based on geophysics aims to enhance the quality of original data, especially for seismic data which is characterized by multiple sources of noise. In fact, the AI-based seismic data denoising is already an independent research field with significant breakthrough and favourable results comparing to traditional seismic denoising methods such as F-K filtering and wavelet thresholding.

3.3.1. Geological facies distribution prediction by CNN

The CNN can be applied for either supervised or unsupervised model construction (de Matos et al., 2007; Laloy et al., 2018).

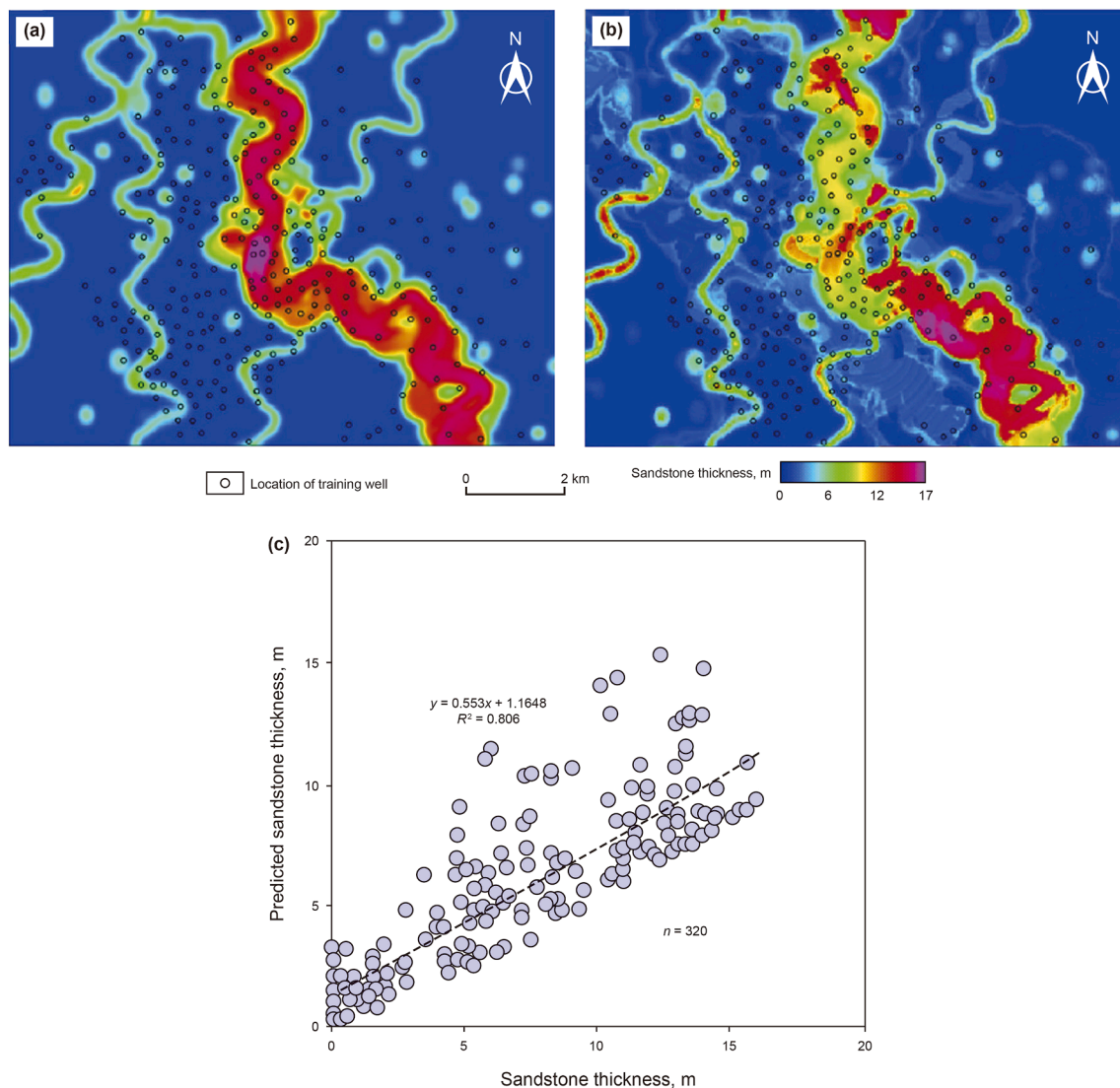


Fig. 14. Sandstone reservoir prediction via SVR (Li et al., 2020). (a) The actual distribution of sandstone reservoir thickness; (b) the predicted distribution of sandstone reservoir thickness via SVR; (c) the correlation between the actual and predicted sandstone thickness.

Supervised CNNs are usually used for petroleum prospect recognition and are based on magnitude-labeled data (Huang et al., 2017; Laloy et al., 2017; Zeiler and Fergus, 2014). As unsupervised algorithms, CNN models with different kinds of autoencoders can create feature maps according to the input seismic images. Through convolution processes without labels, the CNN model is able to learn the features and obtain the information about the characteristics of oil and gas prospects (Long et al., 2015).

(1) Recognition model with a supervised CNN

Most of the original input data for CNNs are labeled seismic images that are processed by eliminating exceptional values and standardizing them to obtain high-quality data (Chevitarese et al., 2018; Zhao, 2018). Then, the convolution of the image allows CNN model to extract the most important features of different geological facies or petroleum reservoirs (Fig. 17(a)). For the training process, the original seismic image with labels of petroleum prospect (facies 2 in Fig. 18) will be convoluted and pooled, during which the feature for each facies is extracted. In this way,

the classification rules are determined to recognize the geological facies and potential targets for hydrocarbon accumulation (Fig. 17(b)–(d)). The CNN model can significantly enhance the geological facies recognizing efficiency and the accuracy comparing to traditional methods based on experiences of professional engineers. At present, the traditional geological facies recognition methods are gradually being replaced by intelligent methods with standard procedure and better performance. While as a supervised algorithm, the CNN model for classification and recognition requires sufficient tagged images (Zeiler and Fergus, 2014). Therefore, the supervised CNN is usually applied in a mature exploration area characterized by enough drilling wells. However, for a district with limited wells and label data, the supervised CNN is often ineffective, and an unsupervised CNN can be applied in this situation.

(2) Prediction model with an unsupervised CNN

The seismic attribute images without labels are processed (usually separated) to meet the requirements of convolution processes (Fig. 18(b)). The input images are subsequently provided to

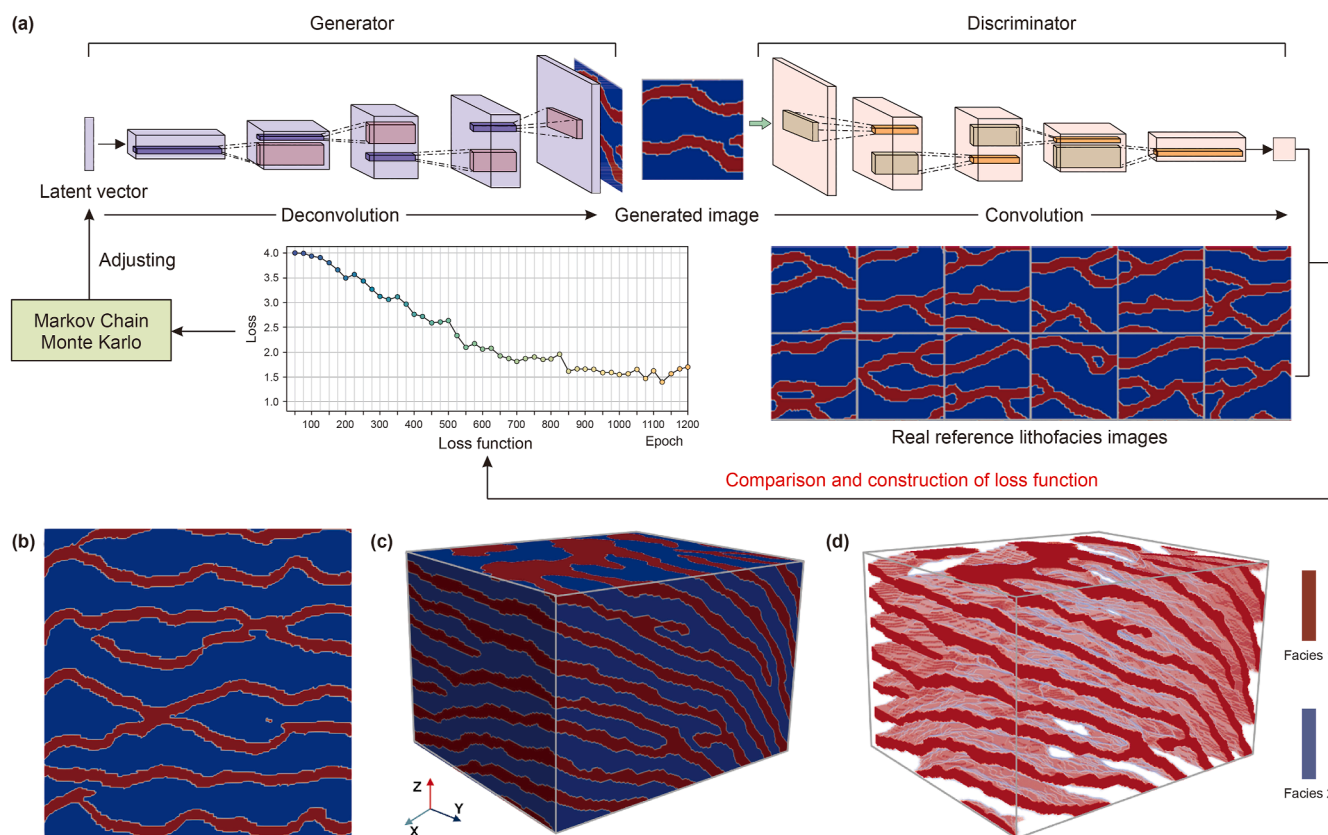


Fig. 15. The structure of GANs for PPEP work in geological facies. (a) The workflow of GANs for geological facies prediction; (b) the 2D facies maps generated by the GAN model; (c) the original 3D geological facies distribution; (d) the prediction of the 3D geological facies distribution shown in (c).

the autoencoder for the CNN model to generate feature maps through a convolutional model. After 4 convolutional processes, the original image is transferred to quantitative information recorded by a latent vector. The deconvolution is exactly the reverse process which will reconstruct the latent vector to a feature map of petroleum reservoir (Fig. 18(a)) (Yue et al., 2019). The convolutional part generates a fully connected layer or an expression layer for the decoder to create a distribution map of geological facies or favourable oil and gas reservoirs, which can be used for the prediction of petroleum prospects. Generally, the output will be transferred to a probability map to indicate the favourable hydrocarbon accumulation region by uncertainty quantification (Fig. 18(c)). (Laloy et al., 2018). The results can also be tested and calibrated with limited drilling data as an example of supervision.

To conclude, CNN is a dominant algorithm for PPEP by geophysical analysis based on 2D and 3D seismic data. The quality of seismic data determines the final performance of CNN models. Therefore, the seismic denoising methods are often set as a general step for the application of CNN in PPEP. The CNN can be directly used for predicting noise by residual neural network, the variational autoencoder (VAE) can learn the distribution of seismic signals and filter noise when reconstructing the seismic data. GANs are also an unsupervised algorithm to denoise by generating noise data and distinguishing them by training the discriminator. Furthermore, active learning achieves noise filter by marking a small part of the key training data with high uncertainty. Such methods can significantly enhance the denoising efficiency (Alfarraj and AlRegib, 2022; Yu et al., 2024).

3.3.2. Tight gas layer prediction by BPNN

As one of the most popular neural networks, the BPNN is usually used for PPEP in reservoir modelling and favourable district recognition by well log data (Jiang et al., 2023; Zhao et al., 2023). There are more than 20 types of well log data and more than 100 attributes of seismic data, which can provide enough data for conducting a DL process via a neural network. Generally, the label of the PPEP model built by the BPNN is the demonstrated favourable space, district or depth range interpreted from well log and seismic data with production responses. Theoretically, a BPNN model can approximate any complex nonlinear continuous function with arbitrary precision due to its powerful nonlinear mapping capability and its self-learning and adaptive ability. However, such a complex nonlinear mapping procedure may lead to intense overfitting and local optima, especially when the data quality is low or noisy, which is unavoidable for the well log data.

There are three main methods to avoid overfitting or local optima for BPNN model: First, the original model is optimized by some advanced optimizers including Momentum, Adam and RMSprop to accelerate convergence and save the training time (Jolliffe, 1986; Schölkopf et al., 1998). Second, activation function can largely mitigate the vanishing gradient problem. Third, the regularization techniques including dropout and L1/L2 regularization are all effective ways for controlling overfitting. The present BPNN model for oil/gas layer identification and prediction for tight gas/shale oil reservoirs is usually composed of 1–2 hidden layers and applies Adam optimizer based on 8–10 types of well log curves (Fig. 19). The predicting accuracy is generally higher than 88% with the training time of about 1–2 GPU hours (Zhao et al., 2023).

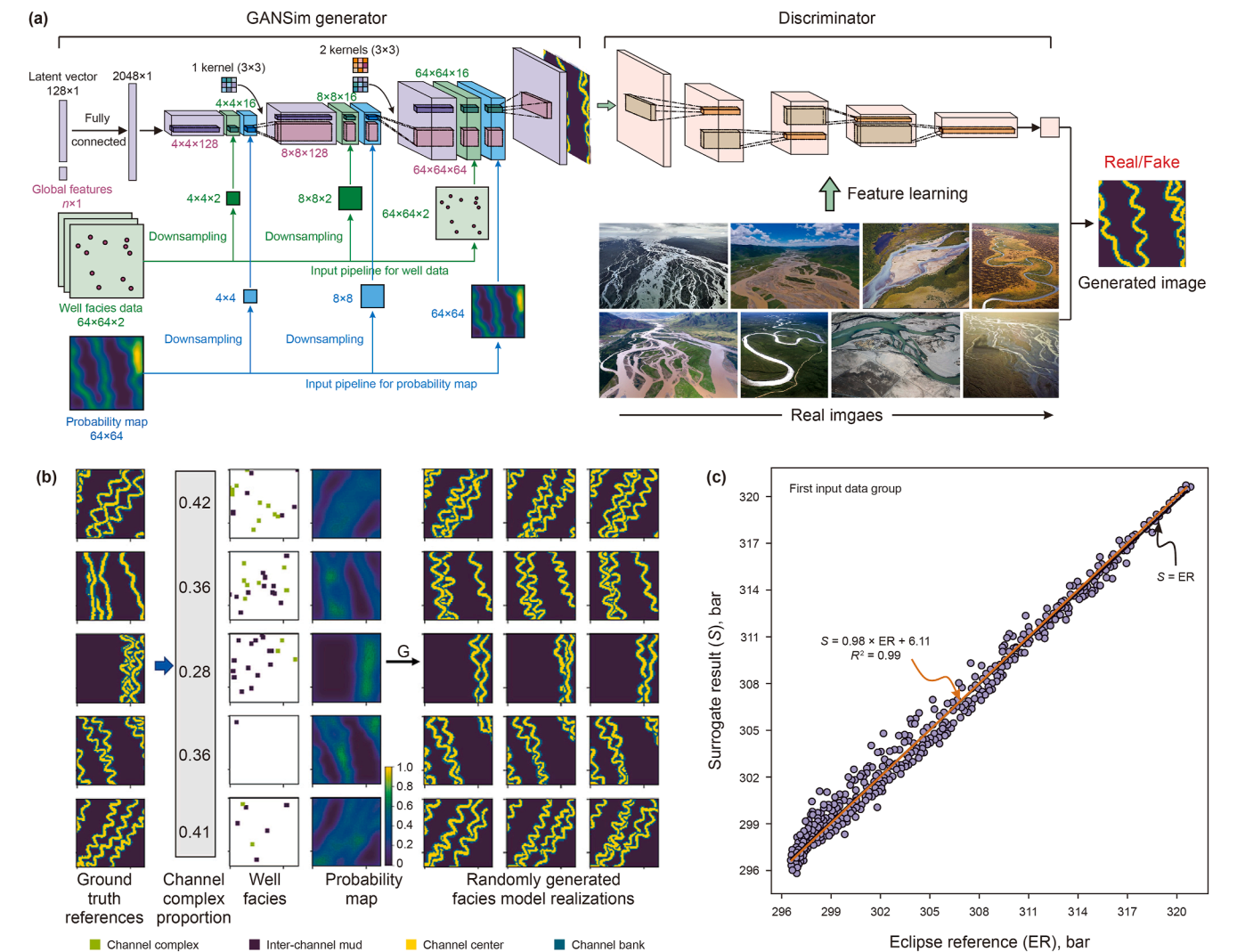


Fig. 16. The favourable carbonate reservoir facies prediction by GANSim (Song et al., 2021). (a) The workflow of GANSim; (b) the distribution of constraint conditions; (c) the prediction results by GANSim.

Table 2
Comparison of ML and DL models applied for PPEP in development geology.

Algorithm	Target	Data type	Data size	Accuracy	Advantage	Limitation
SVR	Sandstone reservoir thickness	Well log; seismic	4×10^5 – 8×10^5	80%–85%	Low cost	Accuracy
GAN	Favourable facies prediction	Seismic	$>10^8$	90%–98%	High accuracy	Robustness; high cost; model collapse
GANSim	Favourable facies prediction	Well log; seismic	$>10^8$	95%–99%	High accuracy; good robustness; low requirement to supervised data	High cost

3.3.3. Hydrocarbon reservoir prediction by transformer

Comparing to BPNN, the transformer architecture solves the local optimum problem through its advanced architecture, which shows impressive robustness and global optimum ability, which is required for the PPEP of unconventional resources that are characterized by intensive heterogeneity (Hua et al., 2024; Lan et al., 2025). In this case, a local optimum can easily lead to errors in predicting duty.

The transformer-based framework requires a large data size and advanced equipment (High-computing-power GPU) (Zhu

et al., 2025). Therefore, the data size and data quality control the performance of the transformer model used in PPEP. The most suitable data for transformer application are multiple types of well log curves with numerous data points. The complete well log data are always split into small segments to save computation. In the case showed in Fig. 20, a total of 10,973 sample points are used for transformer model training to predict the tight gas layer. The multi-head attention mechanism is configured with 5 heads and 3 fully connected layers. The learning rate is set at 0.01, with an iteration count of 500 (Fig. 20(a)) (Hua et al., 2024). The final result

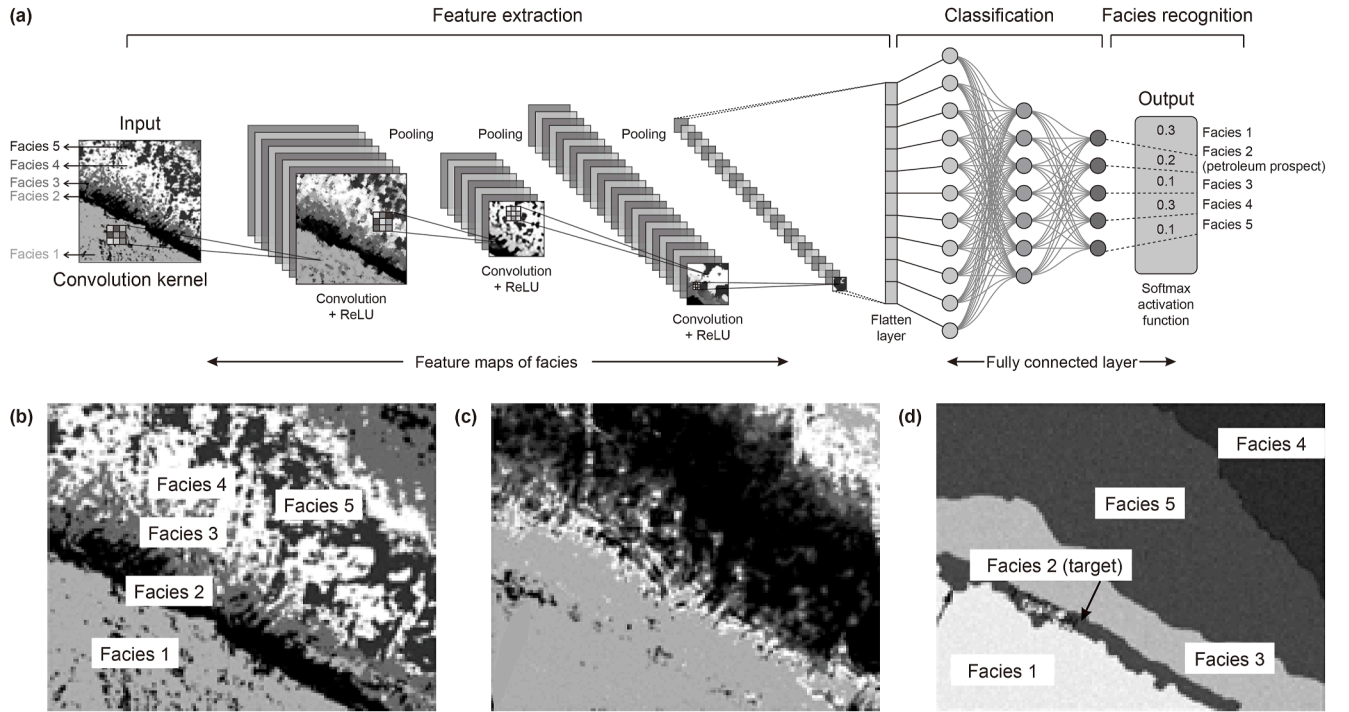


Fig. 17. Favourable facies recognition by a supervised CNN. (a) The geological facies classification and recognition processes of the CNN; (b) the training input seismic image with labels; (c) the testing input image without labels; (d) the classification results of the facies shown in (c) by the CNN model shown in (a) (Chevitarese et al., 2018).

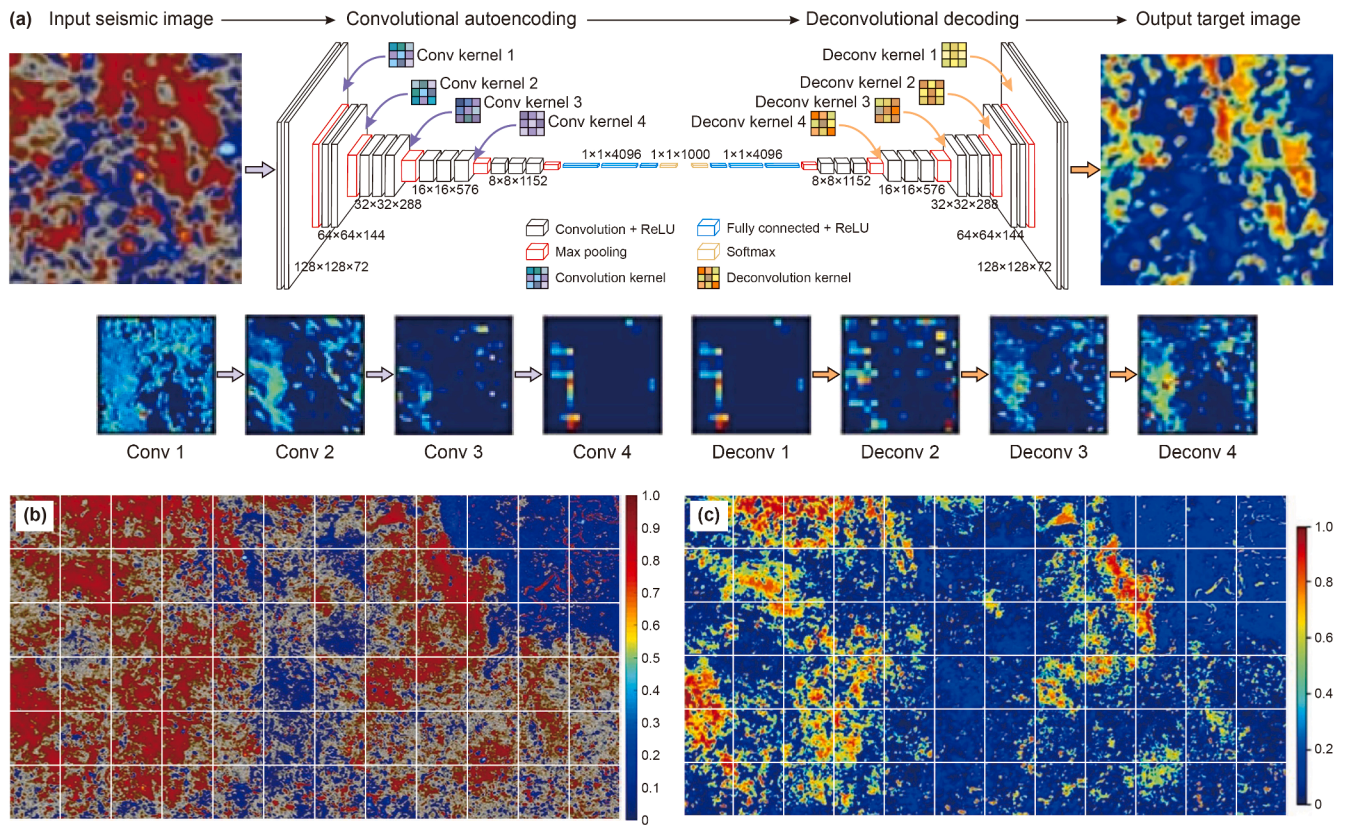


Fig. 18. PPEP for primary sandstone reservoir distribution prediction via an unsupervised CNN. (a) The autoencoding and decoding processes of the CNN; (b) the division of the complete input seismic image to provide a proper image size for convolution; (c) the complete output of the petroleum prospect distribution, which shows the probability of hydrocarbon accumulation (Zhang et al., 2019).

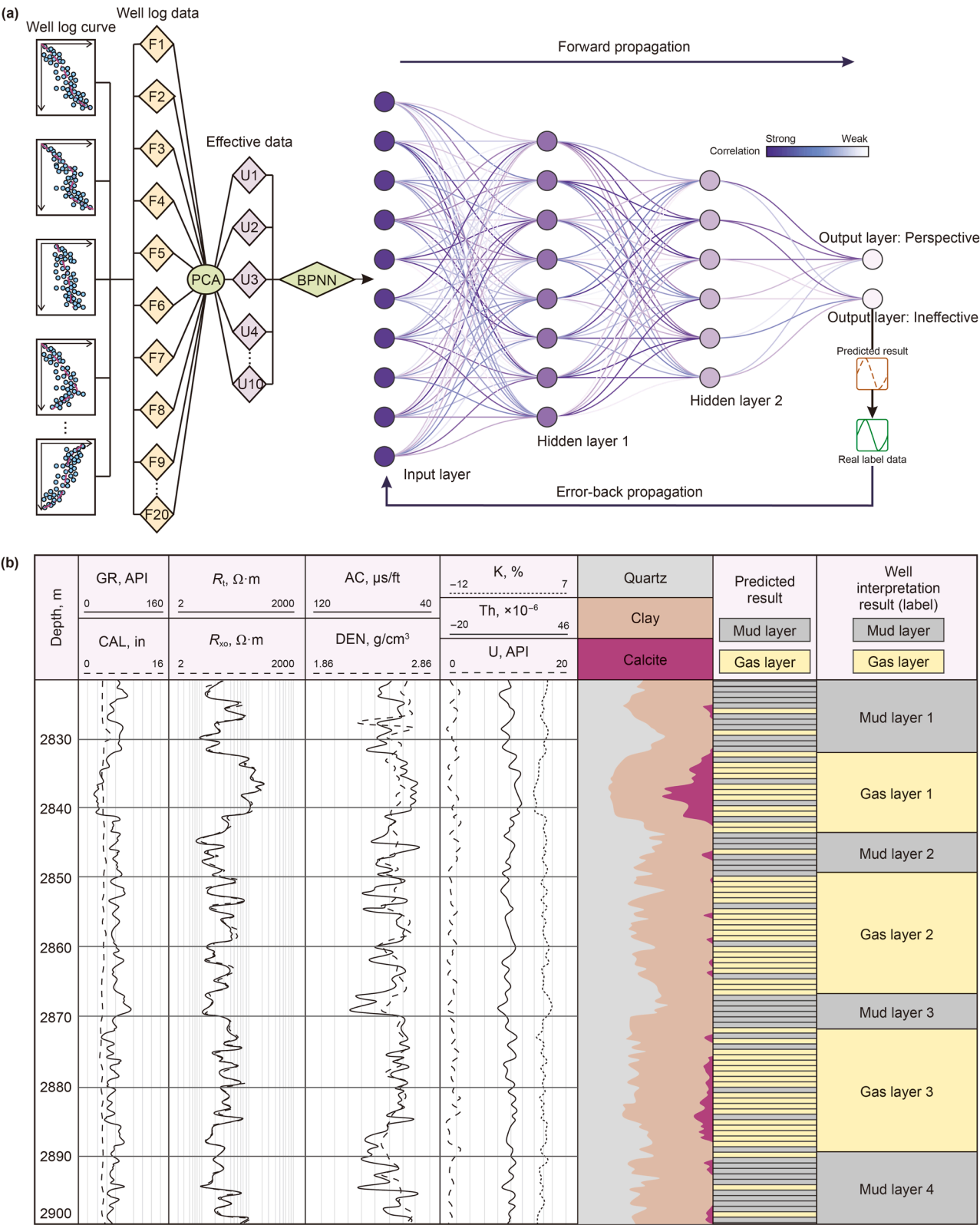


Fig. 19. Gas prospect prediction by the BPNN. (a) Processes of the BPNN model for gas prospect prediction; (b) well log dataset and gas prospect prediction results by the BPNN (Zhao et al., 2023).

shows a prediction accuracy of 94% for gas layer prediction (Fig. 20(b)).

Owing to the positioning operation by attention mechanism, the transformer model always has wider applying scenarios and

better performance. The only requirement of the transformer application is data. For further enhancing the dataset quality, the symbolic aggregate approximation (SAX) method is always used for transferring the well log data to symbolic

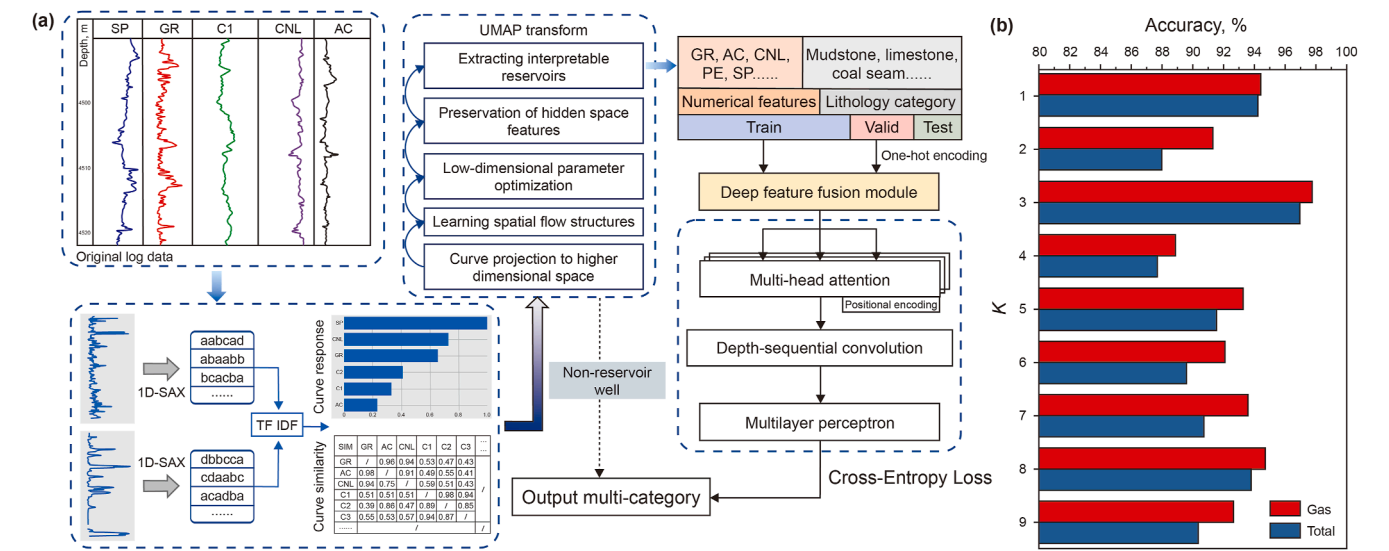


Fig. 20. Tight gas layer prediction by transformer model (Hua et al., 2024). (a) Fluid identification framework based on the transformer; (b) nine-fold cross-validation error bar. It shows identification accuracy for gas layers and all strata classes.

Table 3
Comparison of ML and DL models applied for PPEP in geophysical exploration.

Algorithm	Target	Data type	Data size	Accuracy	Advantage	Limitation
Supervised CNN	Seismic facies recognition	Seismic	$>10^8$	88%–95%	Low cost	Generalization
Unsupervised CNN	Seismic facies prediction	Seismic	$>10^8$	80%–85%	Generalization; unsupervised	Accuracy; geological rationality
BPNN	Tight gas layer prediction	Well log	$>10^6$	85%–90%	Self-optimize	Weak robustness; generalization
Transformer	Lithology recognition; gas/ water layer prediction	Well log	$>10^6$	88%–94%	High accuracy; good robustness; self- optimize	High cost

representations, and then the term frequency-inverse document frequency (TF-IDF) method is used for selecting the best well log curve data.

To conclude, the PPEP in geophysical analysis concentrates on the application of high-quality seismic and well log data. The CNN is suitable for seismic data-based PPEP work owing to the strong image feature extraction ability, while multiple well log data can provide enough data size for the utilization of BPNN and transformer. The denoising strategy of seismic data determines the performance of CNN, while the well log data augmentation also affects the performance of BPNN. Although transformer can fundamentally avoid the model from local optima, it has the highest requirement of data both in size and quality (Table 3).

4. Perspective and future challenges

To conclude, AI methods have been applied to PPEP in many aspects, which has led to good performance. However, there are also ways to further enhance the performance of the models and the practicality of the intelligent method systems for PPEP. In this section, ways of enhancing the accuracy and practicality and future development directions of AI application to PPEP are discussed.

4.1. Accuracy

The accuracy of the ML models is influenced by many factors. Improving the accuracy of an ML model is an extremely complex process. In PPEP, several important factors control model accuracy.

4.1.1. Data (data normalization and data compensation)

Data are the basis of ML and therefore the most important factor for controlling the accuracy of the ML model (Bistline and Merrick, 2020; Breiman, 2001; Huber, 1981; Itani et al., 2020; Le Roux and Bengio, 2008). Generally, the more data we have, the better the model training effect is and the higher the model accuracy is. Unfortunately, in PPEP work, the datasets are relatively small, with hundreds or thousands of data points (except for those based on seismic images). The reason is that the target for PPEP is petroleum prospects or the location of drilling wells. For a petroliferous basin or depression, dozens to hundreds of petroleum reservoirs and drilling wells are as many as possible (Ma et al., 2022; Wang et al., 2024).

However, there are still several ways to overcome the shortage of small data. Here are two main directions:

(1) Enlarging the data size with custom-made data

Self-made data are usually used to train CNNs and GANs in image recognition and feature extraction (Alex et al., 2012; Denton et al., 2015; Szegedy et al., 2015). Some researchers have applied rotation, flipping, and symmetry operations to increase the amount of training data based on limited real images of geological facies and petroleum reservoir bodies, such as rivers, caves and fractures. While the data enlargement has a risk of overfitting, which needs K-fold cross-validation, validation dataset, or learning curve, for validation.

(2) Increasing the quality of the data

Table 4
Classic evaluation parameters in PPEP work based on petroleum geology analysis.

Parameter	Definition	Function	References
Hydrocarbon expulsion rate (HER)	Hydrocarbon quantities produced by source rock per unit area in million years	Evaluating the petroleum and gas provision of the source rocks	Durand (1983); Peters (1986); Tissot and Welte (1978)
Hydrocarbon expulsion intensity (HEI)*	The quantity of hydrocarbon expulsion per unit area	Evaluating the petroleum and gas provision of the source rocks	Collett, 2002
Normalized facies index (NFI)	The total reserve distributed in a certain type of sedimentary facies	Quantifying the sedimentary facies and then reservoirs	Wang et al. (2023)
Hydrocarbon height (HCH)	The height of hydrocarbon accumulation in a trap	Evaluating the abundance of petroleum prospects	Magoon and Dow (1992)
Fluid potential (φ_f)	The total mechanical energy of a unit underground relative to the set base level	Quantifying favourable zones for geofluid migration	England et al. (1987); Hubbert (1953)

Note: $HEI = \int_{z_0}^z 0.1 \times P_e \times H \times \rho \times TOC \times dz$; where z is the burial depth, m; and z_0 is the burial depth of the hydrocarbon expulsion threshold, m, which is determined by the crossplot of $(S_1+S_2)/TOC$ and R_o ; H is the thickness of the source rocks, m; ρ is the density of source rock, g/cm³; and TOC is the total organic content, %.

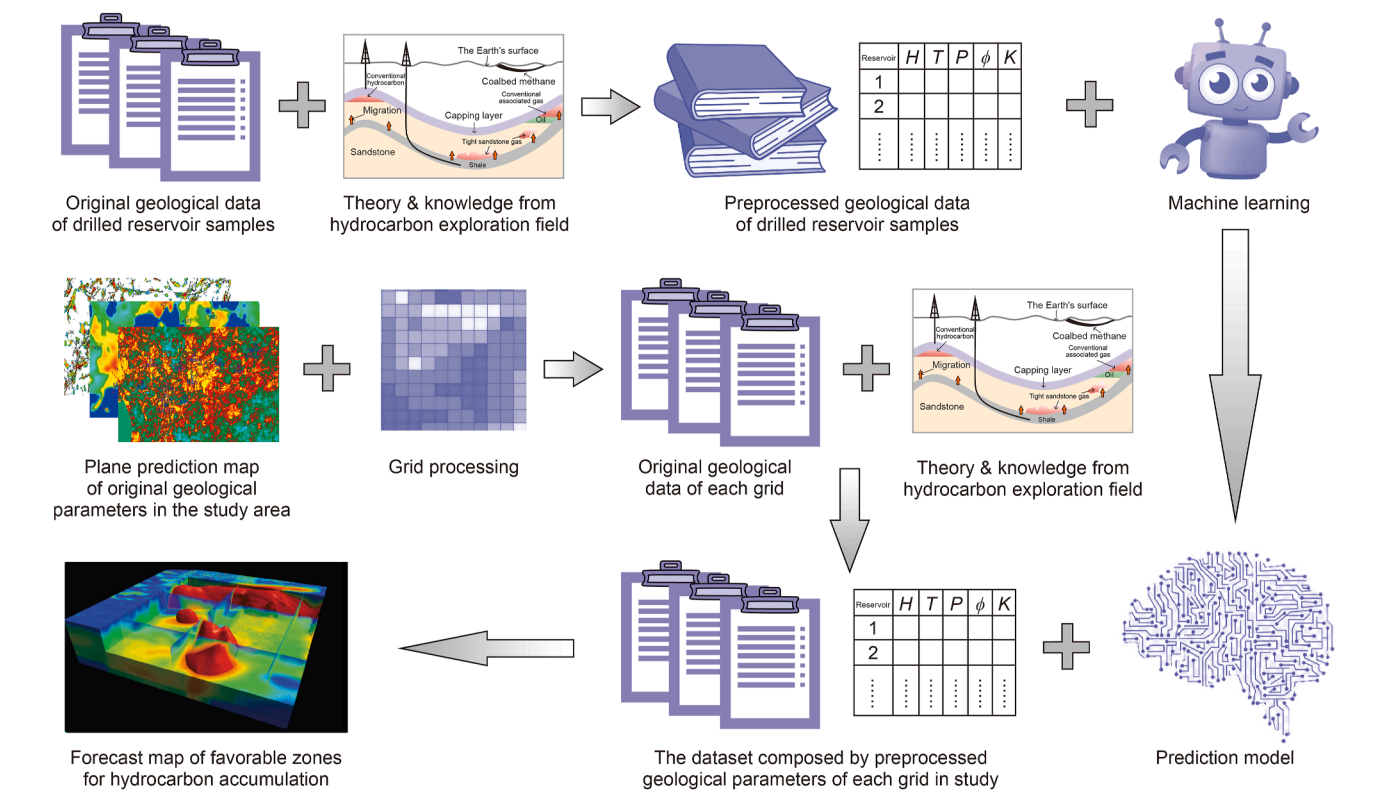


Fig. 21. Dataset construction guided by professional knowledge.

For PPEP work conducted with using most of the general ML methods, the image is not the dominant data type. The best way to overcome the issue of small datasets is to enhance the data quality via professional knowledge. Different kinds of data have different information provision for the ML model. Isolated data can provide only limited data, such as thickness, porosity, permeability etc., whereas some classic and comprehensive evaluation parameters applied in professional fields may contain many types of information (Table 4). With respect to data with professional backgrounds, ML models with small datasets can also achieve good performance, with accuracies greater than 90% (Fig. 21).

(3) Transfer learning

Transfer learning (TL) is a paradigm shift in machine learning wherein knowledge acquired from a source domain is systematically transferred to enhance model performance in a distinct but related target domain (Pan and Yang, 2010). This approach circumvents the conventional requirement for extensive labeled datasets in target tasks, offering significant efficiency gains across computational geoscience, remote sensing, and resource exploration applications (Weiss et al., 2016). For a petroliferous district without efficient drilling and seismic data, the TL can be applied. Based on the professional knowledge extraction and target domain adaption tasks, the TL can generate geological models with high accuracy, enhanced generalization and less computation. Considering the data requirement for the original model and the

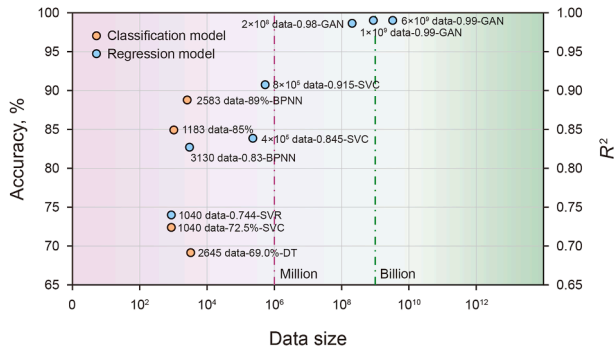


Fig. 22. Accuracy–data size correlation in PPEP work.

similarity of the transferred task, TL is often used for petroleum reservoir characterization and reservoir property prediction. Gao et al. (2024) applied transfer learning framework to predict the tight gas reservoir physical properties in a study area with limited drilling data in Sichuan Basin. The result showed a correlation coefficient of 0.95 between predicted and real data, which indicates the efficiency of TL.

4.1.2. Algorithm

The advantages and disadvantages of existing algorithms have been a topic of constant debate since the birth of AI. However, with decades of development, ML algorithms have been improved by scientific researchers (Alex et al., 2012; Duchi et al., 2011). Previous research has revealed that the priorities of algorithms and the differences between algorithms are evident only if the dataset is large enough (Fig. 22).

For classification problems, models based on DTs, SVMs, and BPNNs performed similarly better than the NB model with dozens or hundreds of data points (Fig. 23(a)). The lower accuracy of the NB model is related to its assumption that the input feature variables are generally independent, which is incorrect because the geological processes are interrelated and interact with each other (Amyx et al., 1960; Gvirtzman and Stanislavsky, 2000; Singer, 1993). For regression problems, the SVR and boosting algorithms showed some advantages (Fig. 23(b)). Notably, this comparison is based on a data size of approximately 1000, which is relatively small for testing the capability of different algorithms. In conclusion, the selection of algorithms is not the key problem in the application of AI in the petroleum industry (the p -value between any two algorithms in accuracy is all lower than 0.05, which indicate the reliability of the comparison results). Instead, the size

and quality of the data should still be the most important areas of focus.

Instead of determining the bottom-level logic, determining the algorithms while considering the characteristics of the data can be more effective. Some researchers have noted the unbalanced distribution of data labels, which is characterized by many more successfully drilled wells and fewer dry (failed) drilled wells; therefore, RUSBoosted decision trees have been applied to enhance model performance (Ma et al., 2022).

4.1.3. Calibration and hyperparameters

The calibration process is very important for model performance, which is achieved mainly by hyperparameter adjustment. The dominant hyperparameters of the different algorithms are listed in Table 5. There are also several methods for hyperparameter determination.

(1) Determination by experience

Experience-based hyperparameter adjustment could be the best or worst method for model calibration. The final results of the calibration depend on the experience of the operators. It is generally applied at the early stage of ML utilization.

(2) Enumeration

Enumeration is the simplest way to search for the best hyperparameter or parameter combination. It begins at the start and ends by trying all the values of the parameters and finding the best one. When there is only one hyperparameter with a limited value range, enumeration can be selected as a method for model calibration. However, when there are several hyperparameters with a wide value range, enumeration to find the best hyperparameter combination may be too time-consuming.

(3) Random search

The principle of random search is similar to enumeration with the only difference being the search order. Therefore, the effect of random search is also unstable because it may find the best hyperparameter(s) on the first attempt or after 100,000 attempts.

(4) Loss function

The loss function is the most common model calibration method and is applied to characterize the relationship between

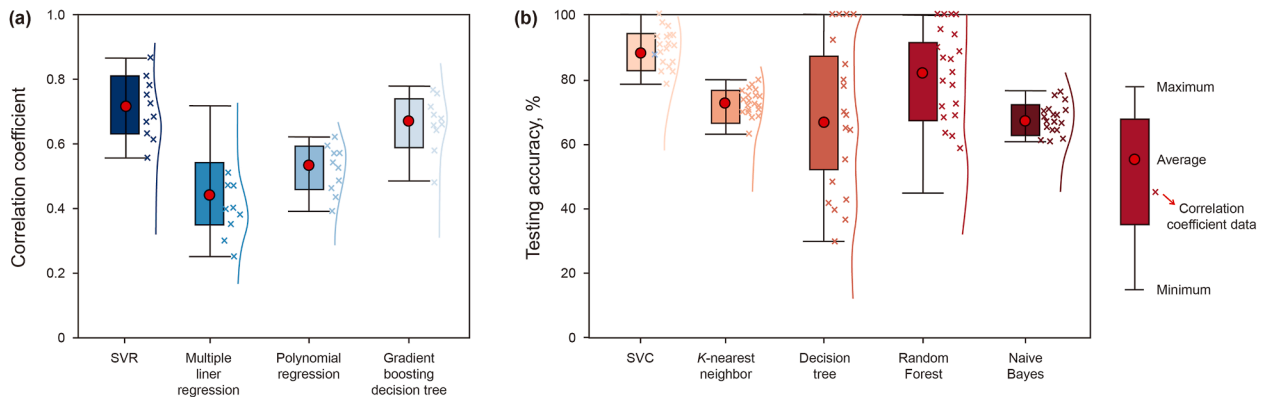


Fig. 23. Comparison of different algorithms used in PPEP work. (a) Accuracy comparison of classification algorithms for PPEP work; (b) accuracy comparison of regression algorithms for PPEP work.

Table 5
Dominant hyperparameters of the ML algorithms for PPEP work.

Algorithm	HP*	Definition/effects	References
DT	Max depth Min samples split Min samples leaf	The maximum layer of a decision tree The minimum samples for splitting The minimum samples in a subgroup	Breiman (2001); Quinlan (1986)
SVM	Kernel γ (Gamma) C-value	Map the input space to a high-dimensional feature space by nonlinear transformation The kernel parameter of RBF The penalty coefficient represents the model tolerance of errors	Burges (1998); Chang and Lin (2001); Cortes and Vapnik (1995); Guyon et al. (1992)
BPNN	Learning rate Batch size Regularization parameters	The stride of weights updating for each iteration Sample size of the iteration The parameters for evaluating the overfitting behaviour and the generalizability of the BP model	Gori and Tesi (1992); Hornik (1991); Robert (1989)
CNN	Hidden layer Node Image pixels Kernel size Kernel amount Stride Padding Activation function	The amount of the hidden layer The number of the node for a certain hidden layer The pixels of the input images The size of the convolutional kernels ($3 \times 3, 5 \times 5$) The amount of the kernel size used for convolution The stride of the convolutional process The added pixels around images during convolution Enhances the expressional ability of a model by nonlinear transformation (ReLU, Sigmoid, Tanh)	Alex et al. (2012); LeCun et al. (1990); Zeiler and Fergus (2014)
GANs	Image size Latent dimension Temporal resolution	The generated image size The dimension of the input latent vector for the generator network The interval for output data	Denton et al. (2015); Goodfellow et al. (2014); Long et al. (2015); Mirza and Osindero (2014)
EL	Training epochs n -estimators Bootstrap	The times or iterations for network training The number of the weak learner in an ensemble model Sampling with replacement or not	Chen et al. (2016); Freund (1999)

Note: HP* represents hyperparameter.

Table 6
Domain loss functions of the ML algorithms for PPEP work.

Loss function	Definition	Application scenarios	References
L1 Loss	The absolute value of the difference between the output and the target	Reservoir prediction	Girshick (2015)
Mean Squared Error (MSE) Loss	The MSE between the output and the target	Logging facies prediction	Köksoy (2006); Liu et al. (2023)
Cross-Entropy Loss KLDiv Loss	The difference between the predicted probability distribution and the real probability distribution	Seismic facies prediction	Lin et al. (2020); Sudre et al. (2017)
Binary Cross-Entropy Loss Soft Margin Loss	The difference between the output and the target in a binary problem	Hydrocarbon accumulation probability prediction	Bosman et al. (2020); Pihlak and Riid (2021)

the predicted and actual values (Girshick, 2015). The hyperparameters are subsequently adjusted with the aim of decreasing the loss, which is the dependent variable of the loss function. The commonly used loss functions for different ML algorithms are listed in Table 6. Generally, L1 Loss and Mean Squared Error (MSE) Loss are widely used for sandstone thickness prediction, and Cross-Entropy Loss and KL Divergence (KLDiv) Loss are applied for geological facies prediction (Köksoy, 2006; Lin et al., 2020; Liu et al., 2023; Sudre et al., 2017). Binary Cross-Entropy Loss and Soft-Margin Loss are used for hydrocarbon accumulation probability prediction models, in which the target of the model usually contains two classes (success or failure) (Bosman et al., 2020; Pihlak and Riid, 2021).

(5) ML optimizers

Besides the hyperparameter and constraints of multiple loss functions, there are also some intelligent ML and DL optimizers applied for ML model optimization. The optimization of ML is usually based on momentum-based optimizers, which include stochastic gradient descent (SGD), Momentum, and Nesterov accelerated gradient (NAG).

The DL models usually use adaptive learning rate optimizers to automatically adjust the learning rate for each parameter, making

them much easier to use and the current default choice. The most popular optimizers are AdaGrad, RMSProp and Adam (adaptive moment estimation). For the selection of optimizers, the SGD, Momentum, or NAG are all effective for ML models usually constructed for PPEP application for hydrocarbon probability prediction; whereas Adam is the best optimizer for DL model based on seismic or well log data in development geology and geophysical applications.

4.2. Generalization and practicality

4.2.1. Workflow standardization

In many fields, the utilization of AI or ML follows a unique workflow, including data preparation and engineering, software, and opening ports for model calibration, which means that the constructed model can be directly used by any unit or individual with authorization. Standardization is important for utilization and improvement in an interdisciplinary context (Duchi et al., 2011). However, in the petroleum exploration field, there is no unified standardization for intelligent PPEP work. The main reasons are listed below:

(1) The data engineering process

There are large disparities in the original data among different oil and gas companies, research institutes and even universities. There are more than 100 data types, including well log data, seismic data and productivity data. In many situations, these data are not compatible with each other, which means that they can be opened and edited only by designated software. Furthermore, the quantification process leads to many different quantification methods based on different understandings and criteria (Ma et al., 2022; Wang et al., 2023).

(2) Professional interpretation software

As stated above, the incompatibility of the original data requires using different professional interpretation software, which will lead to a data gap and inconvenience in the utilization of ML for PPEP.

4.2.2. Generalization and reliability assessment

Considering the complex geological conditions that exist between different petroleum basins or districts, the generalization and reliability of the AI models for PPEP work are significant. While there is no standard. In petroleum geology, the traditional way of prediction is attributed to the predictability of geological facies and geological bodies, whose structure and distribution follow some natural rules described by professional knowledge and visible evidence (Middleton, 1973; Shanmugam, 1997).

For ML and DL methods, the uncertainty analysis for the model is significant for testing the reliability of the model. There are two main types of uncertainty including model-driven uncertainty and data-driven uncertainty. In fact, the uncertainty caused by data may have more influence on the prediction results of the ML model. The data quality enhancement is also the key for AI applications in all fields. For the model-driven uncertainty, there are two popular ways to evaluate it:

- (1) Cross-validation. The K -fold validation is the most suitable way to test the uncertainty of the PPEP models owing to the small data size. The general evaluation metrics (AUC, F1-score) can reflect the reliability of the model. A low variation indicates a robust model that is not overfitting to local spatial structures.
- (2) Ensemble learning. The uncertainty analysis through ensemble learning treats the final model as an ensemble. For the multiple predictions by the model, the mean prediction and the standard deviation at each location are calculated. In brief, a high standard deviation indicates areas where the model's predictions are inconsistent and thus less reliable (high cognitive uncertainty). The model results for these regions require much more work.

While, there is another method for the uncertainty analysis of the PPEP models, which is post-hoc analysis of errors. This procedure is similar to the analysis of dry wells in petroleum industry. The outputs that are wrongly predicted by the PPEP models are analysed by petroleum geological knowledge, which directly links the model performance to geological understanding. Based on the results of feature importance evaluated by the PPEP model, the differences of "exploration decision" between human and machine can be found.

In summary, the uncertainty analysis of the PPEP model can be regarded as a quantitative response to the petroleum exploration degree and an indication of the exploration success. The high confidence of the model reflects more precise data from seismic and drilling, which will provide more geological information for the decision of exploration location and development strategy. In

contrast, the low confidence of the model is attributed to the poor quality of the data, which cannot provide sufficient information for making an accurate decision in PPEP.

4.3. Future challenges and perspective

With the emergence of GPT-4 and Sora, the world has changed. The future will be significantly transformed by super artificial intelligence. There is no doubt that the current utilization of AI in PPEP has some limitations, but with the development of AI in PPEP, traditional patterns of the petroleum industry will also change.

4.3.1. Current limitations of AI in PPEP

Although the application of AI in PPEP has developed fast in recent years, there are also some limitations need to be noticed.

- (1) Data quality. The data applied in PPEP is still of relatively low quality. On the one hand, the data size of PPEP is generally small compared to medicine, transportation and image detection, etc. On the other hand, the industrial property leads to an unbalanced structure for PPEP data, which is characterized by a lack of negative sample (dry well, empty trap and insufficient target, etc.). Furthermore, the precision of the data is uncertain because of multiple origins including outcrop observation, core sample measurement and well log, which is easily influenced by human mind.
- (2) The interpretability of PPEP model is insufficient. Generally, the PPEP models are merely black box models, which makes it difficult to show the calculation or determination processes for the PPEP and hence decreasing the reliability. Also, the limitation of interpretability will increase the difficulty for model optimization.
- (3) The current AI model for PPEP lacks the guidance from geological knowledge. In other words, the combination of AI and traditional petroleum geological framework is still weak. The data-driven AI models can be optimized by embedding domain-specific knowledge to enhance the robustness (Wang et al., 2023; Zhu et al., 2025). We believe that the geological knowledge can provide certain restrictions or guidance for AI model construction and optimization in PPEP. However, such a combination also needs creative ideas and novel ways.

4.3.2. Challenges and chances for AI application in unconventional resources

As mentioned before, the heterogeneity, limited well log data, and multiple sources of noise in seismic data lead to some challenges for AI application in unconventional resources. Although the data augmentation strategy, unsupervised model and multiple denoising AI methods have been applied to deal with these issues, there is still a lot of room for improvement in the current ML and DL models. However, the complexity of unconventional resources and the poor performance of traditional PPEP methods exactly highlight the importance and the opportunities for AI application. Compared to traditional methods, the data-driven ML model and DL model have excellent information mining capabilities from large-scale data. Furthermore, the higher efficiency and lower cost of AI models will greatly accelerate the unconventional resources research compared to expensive laboratory experiments and optical microscopes. To conclude, although there are indeed many challenges currently, it is necessary to develop AI applications in PPEP for unconventional resources.

4.3.3. Future development direction

To improve the performance of PPEP, there are several ways that worth further research. First, a combination of different AI models is a worthwhile way to harness the different advantages of algorithms. Previous studies have shown the potential of combining multiple AI models. Wang et al. (2023) first applied an SVM to predict petroleum prospects and then constructed an SVR model to predict the reserve distribution while restricting the SVC model's results. Fu et al. (2018) inserted an SVM classification model into a geological facies prediction model via a CNN to increase the number of classification labels and greatly improved the prediction performance.

Along with the development and application of transformer models, the multimodal data fusion, which provides a unified framework for text, image and data, will be more and more important. The multimodal fusion can not only effectively expand the data size, but also provide information from different dimensions, which may lead to better performance of AI models.

In addition to data-driven models, the dual-driven models of physics and data are also a development direction of PPEP. Physics-informed neural network (PINN) framework provides a deep neural network architecture that incorporates physics-driven loss functions to constrain and guide the data-driven model, thereby generating results that are physically and geologically consistent. Such framework can not only improve the performance of DL models, but also enhance their interpretability and alleviate potential “black box” issues, which is of great significance in PPEP work.

Furthermore, an artificial general intelligence (AGI) model, which is an AI model with high learning and generalization abilities, will be the final form for petroleum industry. It can also have its own logic and ideas, including its own knowledge map for a region. With a large amount of data, the AGI is able to learn most of the professional knowledge and skills in the petroleum industry and can further apply itself to solve problems that humans cannot solve. While constrained by data quality, computing hardware and industry competition, the accomplishment of AGI needs more time but is expectable.

5. Conclusions

- (1) The application of AI in petroleum prospect evaluation and prediction (PPEP) focuses on petroleum geology, development geology and geophysical exploration through favourable prospect classification and recognition, geological facies prediction and hydrocarbon reservoir prediction. Machine learning methods and deep learning methods are both applied in multiple concrete scenarios without a unique choice or standard procedure. Compared to traditional PPEP methods, ML and DL models show higher accuracy and efficiency in most fields, which is attributed to the quick development of AI industry in workflow creation and algorithm optimization. With their excellent data processing and self-optimization ability, the AI application will play an increasingly important role in PPEP works.
- (2) The dominant limitations for utilization of AI method in PPEP and petroleum industry are data quality and model interpretability. Data size and data precision are mainly constrained by limited drilling wells and magnitude of seismic noise. Although multiple data augmentation methods, ensemble learning and transfer learning are helpful for enhancing the accuracy of ML or DL models, they do not solve the data-driven problem fundamentally. Model interpretability is significant for the reliability evaluation. In

fact, the interpretability of AI model determines the acceptance of AI technology in the petroleum industry, especially for the petroleum geology field in which PPEP works requires abundant theoretical basis and rigorous logical processes.

- (3) At present, the physics- and data-driven dual models and multimodal fusion technologies are important research directions for the intelligence and digitization of the oil and gas industry, which may greatly enhance data quality and model interpretability. Furthermore, the AGI model, which has high learning and generalization abilities, should be the final goal for the application of AI in the petroleum industry.

CRedit authorship contribution statement

Qiao-Chu Wang: Writing – review & editing, Writing – original draft, Conceptualization. **Dong-Xia Chen:** Writing – review & editing, Data curation, Conceptualization. **Sha Li:** Investigation, Conceptualization. **Yu-He Chen:** Investigation, Data curation. **Xiang Chen:** Data curation. **Mao-Sen Wang:** Writing – review & editing. **Zai-Quan Yang:** Validation, Data curation. **Fu-Wei Wang:** Methodology, Conceptualization.

Declaration of competing interest

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

Acknowledgements

This study was supported by the National Natural Science Foundation of China (Grant No. 42302141). We gratefully acknowledge the Shengli Oilfield Branch of the China Petroleum & Chemical Corporation for providing field test data and reservoir data.

References

- Adnan, M.S., Safaei-Farouji, M., Amiri-Ramsheh, B., et al., 2024. Modeling apparent viscosity of waxy crude oils doped with polymers using tree-based models, radial basis function neural networks, and Gaussian process regression. *Geoenvironment Sci. Eng.* 235, 212689. <https://doi.org/10.1016/j.geoen.2024.212689>.
- Agterberg, F., 2011. A modified weights-of-evidence method for regional mineral resource estimation. *Nat. Resour. Res.* 20 (2), 95–101. <https://doi.org/10.1007/s11053-011-9138-0>.
- Akkaş, E., Akin, L., Evren Çubukçu, H., et al., 2015. Application of decision tree Algorithm for classification and identification of natural minerals using SEM-EDS. *Comput. Geosci.* 80, 38–48. <https://doi.org/10.1016/j.cageo.2015.03.015>.
- Al-Anazi, A.F., Gates, I.D., 2010. Support vector regression for porosity prediction in a heterogeneous reservoir: A comparative study. *Comput. Geosci.* 36 (12), 1494–1503. <https://doi.org/10.1016/j.cageo.2010.03.022>.
- Alex, K., Ilya, S., Geoffrey, E.H., 2012. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* 25, 2.
- Alfarraj, M., AlRegib, G., 2022. Physics-guided deep learning for seismic interpretation. *Geophysics* 87 (1), 15–28. <https://doi.org/10.1190/geo2020-0312.1>.
- Allan, U.S., 1989. Model for hydrocarbon migration and entrapment within faulted structures. *AAPG Bull.* 73 (7), 803–811. <https://doi.org/10.1306/44b4a271-170a-11d7-8645000102c1865d>.
- Amyx, J.W., Bass, D.M., Whiting, R.L., 1960. *Petroleum Reservoir Engineering, Physical Properties*. McGraw Hill, New York.
- Ankerst, M., Breunig, M.M., Kriegel, H.P., et al., 1999. OPTICS: Ordering points to identify the clustering structure. In: *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*. ACM, pp. 49–60. <https://doi.org/10.1145/304182.304187>.
- Armitage, D.A., Stright, L., 2010. Modeling and interpreting the seismic-reflection expression of sandstone in an ancient mass-transport deposit dominated deep-water slope environment. *Mar. Petrol. Geol.* 27 (1), 1–12. <https://doi.org/10.1016/j.marpetgeo.2009.08.013>.

- Avalos, S., Ortiz, J.M., 2019. Recursive convolutional neural networks in a multiple-point statistics framework. In: *Mining Goes Digital*. CRC Press, London, pp. 168–176. <https://doi.org/10.1201/9780429320774-19>.
- Bashkova, S.E., Kozhevnikova, E.E., Yakovlev, Y.A., et al., 2023. The deep structure of the Proterozoic complex of the sedimentary cover the east European platform: Oil and gas potential. *Geotectonics* 57 (4), 428–445. <https://doi.org/10.1134/s0016852123040039>.
- Bengio, Y., Simard, P., Frasconi, P., 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Network.* 5 (2), 157–166. <https://doi.org/10.1109/72.279181>.
- Bennett, K.P., Mangasarian, O.L., 1992. Robust linear programming discrimination of two linearly inseparable sets. *Optim. Methods Software* 1 (1), 23–34. <https://doi.org/10.1080/10556789208805504>.
- Bistline, J.E.T., Merrick, J.H., 2020. Parameterizing open-source energy models: Statistical learning to estimate unknown power plant attributes. *Appl. Energy* 269, 114941. <https://doi.org/10.1016/j.apenergy.2020.114941>.
- Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. ACM, pp. 144–152. <https://doi.org/10.1145/130385.130401>.
- Bosman, A.S., Engelbrecht, A., Helbig, M., 2020. Visualising basins of attraction for the cross-entropy and the squared error neural network loss functions. *Neurocomputing* 400, 113–136. <https://doi.org/10.1016/j.neucom.2020.02.113>.
- Breiman, L., 2001. Random forests. *Mach. Learn.* 45 (1), 5–32. <https://doi.org/10.1023/A:1010933404324>.
- Breiman, L., Friedman, J., Olshen, R., et al., 1984. *Classification and Regression Trees*. Wadsworth, California.
- Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* 2 (2), 121–167. <https://doi.org/10.1023/A:1009715923555>.
- Chan, P.P.K., Zheng, J., Liu, H., et al., 2021. Robustness analysis of classical and fuzzy decision trees under adversarial evasion attack. *Appl. Soft Comput.* 107, 107311. <https://doi.org/10.1016/j.asoc.2021.107311>.
- Chang, C.C., Lin, C.J., 2001. Training v-support vector classifiers: Theory and algorithms. *Neural Comput.* 13 (9), 2119–2147. <https://doi.org/10.1162/089976601750399335>.
- Charles, E., 2003. Using the triangle inequality to accelerate K-means. In: *International Conference on Machine Learning*, pp. 147–153.
- Chen, X., Duan, Y., Houthoofd, R., et al., 2016. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. *Advances in neural information processing systems*, 29, arXiv:1606.03657 [cs.LG]. <https://doi.org/10.48550/arXiv.1606.03657>.
- Chen, Z., Osadetz, K.G., 2006. Geological risk mapping and prospect evaluation using multivariate and Bayesian statistical methods, western Sverdrup Basin of Canada. *AAPG Bull.* 90 (6), 859–872. <https://doi.org/10.1306/01160605050>.
- Cheng, Y., 1995. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* 17 (8), 790–799. <https://doi.org/10.1109/34.400568>.
- Chevitarese, D., Szwarcman, D., Silva, R.M.D., et al., 2018. Seismic facies segmentation using deep learning. In: *2018 AAPG Annual Convention & Exhibition*. American Association of Petroleum Geologists, 42286. <https://doi.org/10.1306/42286chevitarese2018>.
- Choromanska, A., Henaff, M., Mathieu, M., et al., 2014. The loss surfaces of multilayer networks. *Eprint Arxiv* 192–204. <https://doi.org/10.48550/arXiv.1412.0233>.
- Coimbra, R., Rodriguez-Galiano, V., Olóriz, F., et al., 2014. Regression trees for modeling geochemical data: An application to Late Jurassic carbonates (Ammonitico Rosso). *Comput. Geosci.* 73, 198–207. <https://doi.org/10.1016/j.cageo.2014.09.007>.
- Collett, T.S., 2002. Energy resource potential of natural gas hydrates. *AAPG Bull.* 86 (11), 1971–1992. <https://doi.org/10.1306/61eadd2-173e-11d7-8645000102c1865d>.
- Comaniciu, D., Meer, P., 2002. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 24 (5), 603–619. <https://doi.org/10.1109/34.1000236>.
- Cortes, C., Vapnik, V., 1995. Support-vector networks. *Mach. Learn.* 20 (3), 273–297. <https://doi.org/10.1007/BF00994018>.
- Cox, D.P., 1993. Estimation of undiscovered deposits in quantitative mineral resource assessments: Examples from Venezuela and Puerto Rico. *Nonrenewable Resour.* 2 (2), 82–91. <https://doi.org/10.1007/BF02272805>.
- de Matos, M.C., Osorio, P.L.M., Johann, P.R.S., 2007. Unsupervised seismic facies analysis using wavelet transform and self-organizing maps. *Geophysics* 72 (1), P9–P21. <https://doi.org/10.1190/1.2392789>.
- Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc., Ser. B Stat. Methodol.* 39 (1), 1–22. <https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>.
- Demyanov, V., Backhouse, L., Christie, M., 2015. Geological feature selection in reservoir modelling and history matching with multiple Kernel learning. *Comput. Geosci.* 85, 16–25. <https://doi.org/10.1016/j.cageo.2015.07.014>.
- Denton, E., Denton, E., Chintala, S., et al., 2015. Deep generative image models using a Laplacian pyramid of adversarial networks. In: *Proceedings of the 29th International Conference on Neural Information Processing Systems*, Volume 1. ACM, pp. 1486–1494. <https://doi.org/10.5555/2969239.2969405>.
- Dorotic, M., Stagno, E., Warlop, L., 2024. AI on the street: Context-dependent responses to artificial intelligence. *Int. J. Res. Market.* 41 (1), 113–137. <https://doi.org/10.1016/j.ijresmar.2023.08.010>.
- Dow, W.G., 1974. Application of oil-correlation and source-rock data to exploration in Williston Basin. *AAPG Bull.* 58 (7), 1253–1262. <https://doi.org/10.1306/83d91655-16c7-11d7-8645000102c1865d>.
- Drucker, H., Burges, C., Kaufman, L., et al., 1997. Support vector regression machines. *Adv. Neural Inf. Process. Syst.* 9, 155–161.
- Duchi, J.C., Hazan, E., Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* 12, 2121–2159.
- Dunn, J.C., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J. Cybern.* 3 (3), 32–57. <https://doi.org/10.1080/01969727308546046>.
- Durand, B., 1983. *Advances in Organic Chemistry*. John Wiley, New York.
- England, W.A., MacKenzie, A.S., Mann, D.M., et al., 1987. The movement and entrapment of petroleum fluids in the subsurface. *J. Geol. Soc.* 144 (2), 327–347. <https://doi.org/10.1144/gsjgs.144.2.0327>.
- Fang, J., Yan, Z., Lu, X., et al., 2024. An oil production prediction approach based on variational mode decomposition and ensemble learning model. *Comput. Geosci.* 193, 105734. <https://doi.org/10.1016/j.cageo.2024.105734>.
- Freund, Y., 1995. Boosting a weak learning algorithm by majority. *Inf. Comput.* 121 (2), 256–285. <https://doi.org/10.1006/inco.1995.1136>.
- Freund, Y., 1999. An adaptive version of the boost by majority algorithm. In: *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*. ACM, pp. 102–113. <https://doi.org/10.1145/307400.307419>.
- Friedman, J., Hastie, T., Tibshirani, R., 2000. Special invited paper. Additive logistic regression: A statistical view of boosting. *Ann. Stat.* 28 (2), 337–374.
- Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* 29 (5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>.
- Friedman, N., Geiger, D., Goldszmidt, M., 1997. Bayesian network classifiers. *Mach. Learn.* 29 (2–3), 131–163. <https://doi.org/10.1023/a:1007465528199>.
- Fu, C., Lin, N.T., Zhang, D., et al., 2018. Prediction of reservoirs using multi-component seismic data and the deep learning method. *Chin. J. Geophys.* 61 (1), 293–303 (in Chinese). <https://doi.org/10.6038/cjg2018L0193>.
- Gafni, R., Aviv, I., Kantsepolsky, B., et al., 2024. Objectivity by design: The impact of AI-driven approach on employees' soft skills evaluation. *Inf. Software Technol.* 170, 107430. <https://doi.org/10.1016/j.infsof.2024.107430>.
- Gao, Z., Bai, L., Hu, Q., et al., 2024. Shale oil migration across multiple scales: A review of characterization methods and different patterns. *Earth Sci. Rev.* 254, 104819. <https://doi.org/10.1016/j.earscirev.2024.104819>.
- Girshick, R., 2015. Fast R-CNN. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, pp. 1440–1448. <https://doi.org/10.1109/iccv.2015.169>.
- Gjelsvik, E.L., Fossen, M., Tøndel, K., 2023. Current overview and way forward for the use of machine learning in the field of petroleum gas hydrates. *Fuel* 334, 126696. <https://doi.org/10.1016/j.fuel.2022.126696>.
- Goodfellow, I., Pouget, A.J., Mirza, M., et al., 2014. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* 27, 2672–2680.
- Gori, M., Tesi, A., 1992. On the problem of local minima in backpropagation. *IEEE Trans. Pattern Anal. Mach. Intell.* 14 (1), 76–86.
- Guyon, I., Betozer, B., Vapnik, V., et al., 1992. Automatic capacity tuning of very large VC-dimension classifiers. In: *Proceedings of the 6th International Conference on Neural Information Processing Systems*. ACM, pp. 147–155. <https://doi.org/10.5555/2987061.2987080>.
- Gvirtzman, H., Stanislavsky, E., 2000. Palaeohydrology of hydrocarbon maturation, migration and accumulation in the Dead Sea rift. *Basin Res.* 12 (1), 79–93. <https://doi.org/10.1046/j.1365-2117.2000.00111.x>.
- Han, D., Jung, J., Kwon, S., 2020. Comparative study on supervised learning models for productivity forecasting of shale reservoirs based on a data-driven approach. *Appl. Sci.* 10 (4), 1267. <https://doi.org/10.3390/app10041267>.
- Hastie, T., Tibshirani, R., Friedman, J., 2008. *The Elements of Statistical Learning*, second ed. Springer, Berlin.
- Hinton, G.E., Krizhevsky, A., Wang, S.D., 2011. Transforming auto-encoders. In: *Artificial Neural Networks and Machine Learning – ICANN 2011*. Springer, Berlin, Heidelberg, pp. 44–51. https://doi.org/10.1007/978-3-642-21735-7_6.
- Hinton, G.E., Osindero, S., Teh, Y.W., 2006. A fast learning algorithm for deep belief nets. *Neural Comput.* 18 (7), 1527–1554. <https://doi.org/10.1162/neco.2006.18.7.1527>.
- Hornik, K., 1991. Approximation capabilities of multilayer feedforward networks. *Neural Netw.* 4 (2), 251–257. [https://doi.org/10.1016/0893-6080\(91\)90009-T](https://doi.org/10.1016/0893-6080(91)90009-T).
- Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. *Neural Netw.* 2 (5), 359–366. [https://doi.org/10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8).
- Hornik, K., Stinchcombe, M., White, H., 1990. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Netw.* 3 (5), 551–560. [https://doi.org/10.1016/0893-6080\(90\)90005-6](https://doi.org/10.1016/0893-6080(90)90005-6).
- Hua, Y., Gao, G., He, D., et al., 2024. Reservoir fluid identification based on multi-head attention with UMAP. *Geoenergy Sci. Eng.* 238, 212888. <https://doi.org/10.1016/j.geoen.2024.212888>.
- Huang, L., Dong, X., Clee, T.E., 2017. A scalable deep learning platform for identifying geologic features from seismic attributes. *Lead. Edge* 36 (3), 249–256. <https://doi.org/10.1190/le36030249.1>.
- Hubbert, M.K., 1953. Entrapment of petroleum under hydrodynamic conditions. *AAPG Bull.* 37 (8), 1954–2026. <https://doi.org/10.1306/5ceadd61-16bb-11d7-8645000102c1865d>.
- Huber, P.J., 1981. *Robust Statistics*. John Wiley and Sons, New York. <https://doi.org/10.1002/0471725250>.

- Huo, Y., Bouffard, F., Joós, G., 2021. Decision tree-based optimization for flexibility management for sustainable energy microgrids. *Appl. Energy* 290, 116772. <https://doi.org/10.1016/j.apenergy.2021.116772>.
- Imamverdiyev, Y., Sukhostat, L., 2019. Lithological facies classification using deep convolutional neural network. *J. Petrol. Sci. Eng.* 174, 216–228. <https://doi.org/10.1016/j.petrol.2018.11.023>.
- Itani, S., Lecron, F., Fortemps, P., 2020. A one-class classification decision tree based on kernel density estimation. *Appl. Soft Comput.* 91, 106250. <https://doi.org/10.1016/j.asoc.2020.106250>.
- Jensen, F.V., 1996. *An Introduction to Bayesian Networks*. Springer Verlag, New York.
- Jiang, D., Chen, H., Xing, J., et al., 2023. A new method for dynamic predicting porosity and permeability of low permeability and tight reservoir under effective overburden pressure based on BP neural network. *Geoenery Sci. Eng.* 226, 211721. <https://doi.org/10.1016/j.geoen.2023.211721>.
- Jolliffe, I.T., 1986. *Principal Component Analysis*. Springer, New York, New York. <https://doi.org/10.1007/978-1-4757-1904-8>.
- Kam, T., 1995. Random decision forests. In: *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, pp. 278–282.
- Kamiński, B., Jakubczyk, M., Szufel, P., 2018. A framework for sensitivity analysis of decision trees. *Cent. Eur. J. Oper. Res.* 26 (1), 135–159. <https://doi.org/10.1007/s10100-017-0479-6>.
- Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., et al., 2001. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Comput.* 13 (3), 637–649. <https://doi.org/10.1162/089976601300014493>.
- Kelter, R., 2021. Bayesian model selection in the M-open setting: Approximate posterior inference and subsampling for efficient large-scale leave-one-out cross-validation via the difference estimator. *J. Math. Psychol.* 100, 102474. <https://doi.org/10.1016/j.jmp.2020.102474>.
- Kim, K., 2016. A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree. *Pattern Recogn.* 60, 157–163. <https://doi.org/10.1016/j.patcog.2016.04.016>.
- Köksay, O., 2006. Multiresponse robust design: Mean square error (MSE) criterion. *Appl. Math. Comput.* 175 (2), 1716–1729. <https://doi.org/10.1016/j.amc.2005.09.016>.
- Laloy, E., Héralut, R., Jacques, D., et al., 2018. Training-image based geostatistical inversion using a spatial generative adversarial neural network. *Water Resour. Res.* 54 (1), 381–406. <https://doi.org/10.1002/2017wr022148>.
- Laloy, E., Héralut, R., Lee, J., et al., 2017. Inversion using a new low-dimensional representation of complex binary geological media based on a deep neural network. *Adv. Water Resour.* 110, 387–405. <https://doi.org/10.1016/j.advwatres.2017.09.029>.
- Lan, W., Tao, Y., Liang, B., et al., 2025. Shear wave velocity prediction based on Bayesian-optimized multi-head attention mechanism and CNN-BiLSTM. *Comput. Geosci.* 195, 105787. <https://doi.org/10.1016/j.cageo.2024.105787>.
- Lawson, E., Smith, D., Sofge, D., et al., 2017. Decision forests for machine learning classification of large, noisy seafloor feature sets. *Comput. Geosci.* 99, 116–124. <https://doi.org/10.1016/j.cageo.2016.10.013>.
- Le Roux, N., Bengio, Y., 2008. Representational power of restricted Boltzmann machines and deep belief networks. *Neural Comput.* 20 (6), 1631–1649. <https://doi.org/10.1162/neco.2008.04-07-510>.
- LeCun, Y., Boser, B., Denker, J.S., et al., 1989. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1 (4), 541–551. <https://doi.org/10.1162/neco.1989.1.4.541>.
- LeCun, Y., Boser, B., Denker, J.S., et al., 1990. Handwritten digit recognition with a back-propagation network. *Adv. Neural Inf. Process. Syst.* 2 (2), 396–404.
- LeCun, Y., Bottou, L., Bengio, Y., et al., 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86 (11), 2278–2324. <https://doi.org/10.1109/5.726791>.
- Lee, P.J., Wang, P.C.C., 1983. Probabilistic formulation of a method for the evaluation of petroleum resources. *J. Int. Assoc. Math. Geol.* 15 (1), 163–181. <https://doi.org/10.1007/BF01030081>.
- Lee, Y.J., Mangasarian, O.L., 2001. SSVM: A smooth support vector machine for classification. *Comput. Optim. Appl.* 20 (1), 5–22. <https://doi.org/10.1023/A:101215321374>.
- Levorsen, A.L., 1956. *Geology of Petroleum*. W. H. Freeman and Company, San Francisco.
- Li, W., Yue, D., Wu, S., et al., 2020. Thickness prediction for high-resolution stratigraphic interpretation by fusing seismic attributes of target and neighboring zones with an SVR algorithm. *Mar. Petrol. Geol.* 113, 104153. <https://doi.org/10.1016/j.marpetgeo.2019.104153>.
- Lin, T.Y., Goyal, P., Girshick, R., et al., 2020. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 42 (2), 318–327. <https://doi.org/10.1109/TPAMI.2018.2858826>.
- Linde, N., Renard, P., Mukerji, T., et al., 2015. Geological realism in hydrogeological and geophysical inverse modeling: A review. *Adv. Water Resour.* 86, 86–101. <https://doi.org/10.1016/j.advwatres.2015.09.019>.
- Liu, Q., Wang, Z., He, X., et al., 2023. An approximate minimum mean-square error estimator for linear discrete time-varying systems: Handling try-once-discard protocol. *Automatica* 147, 110656. <https://doi.org/10.1016/j.automatica.2022.110656>.
- Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 3431–3440. <https://doi.org/10.1109/CVPR.2015.7298965>.
- Ma, K., Pang, X., Pang, H., et al., 2022. A novel method for favorable zone prediction of conventional hydrocarbon accumulations based on RUSBoosted tree machine learning algorithm. *Appl. Energy* 326, 119983. <https://doi.org/10.1016/j.apenergy.2022.119983>.
- MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. In: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.
- Magoon, L.B., Dow, W.G., 1992. The petroleum system-status of research and methods. *USGS Bulletin* 20 (7), 98.
- Martin, E., Hanspeter, K., Jorg, S., et al., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231.
- Martinelli, G., Eidsvik, J., Hauge, R., et al., 2011. Bayesian networks for prospect analysis in the North Sea. *AAPG Bull.* 95 (8), 1423–1442. <https://doi.org/10.1306/01031110110>.
- Middleton, G.V., 1973. Johannes Walther's law of the correlation of facies. *Geol. Soc. Am. Bull.* 84 (3), 979. [https://doi.org/10.1130/0016-7606\(1973\)84<979:jwlot>2.0.co;2](https://doi.org/10.1130/0016-7606(1973)84<979:jwlot>2.0.co;2).
- Miletić, M., Küçüksuysal, C., Gülcan, M., et al., 2024. Imaging porosity evolution of tight sandstone during spontaneous water imbibition by X-ray Micro-CT. *Heliyon* 10 (11), e31844. <https://doi.org/10.1016/j.heliyon.2024.e31844>.
- Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. *Comput. Sci.* 2672–2680. <https://doi.org/10.48550/arXiv.1411.1784>.
- Mnih, V., Kavukcuoglu, K., Silver, D., et al., 2015. Human-level control through deep reinforcement learning. *Nature* 518 (7540), 529–533. <https://doi.org/10.1038/nature14236>.
- Pan, S.J., Yang, Q., 2010. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22 (10), 1345–1359. <https://doi.org/10.1109/TKDE.2009.191>.
- Perrodon, A., Masse, P., 1984. Subsidence, sedimentation and petroleum systems. *J. Petrol. Geol.* 7 (1), 5–25. <https://doi.org/10.1111/j.1747-5457.1984.tb00158.x>.
- Peters, K.E., 1986. Guidelines for evaluating petroleum source rock using programmed pyrolysis. *AAPG Bull.* 70 (3), 318–329. <https://doi.org/10.1306/94885688-1704-11d7-8645000102c1865d>.
- Pihlak, R., Riid, A., 2021. Morphological cross entropy loss for improved semantic segmentation of small and thin objects. *Procedia Comput. Sci.* 192, 582–591. <https://doi.org/10.1016/j.procs.2021.08.060>.
- Platt, J.C., 1998. Fast training of support vector machines using sequential minimal optimization. In: *Advances in Kernel Methods*. The MIT Press, Cambridge. <https://doi.org/10.7551/mitpress/1130.003.0016>.
- Powley, D.E., 1990. Pressures and hydrogeology in petroleum basins. *Earth Sci. Rev.* 29 (1–4), 215–226. [https://doi.org/10.1016/0012-8252\(0\)90038-W](https://doi.org/10.1016/0012-8252(0)90038-W).
- Pradhan, B., 2013. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. *Comput. Geosci.* 51, 350–365. <https://doi.org/10.1016/j.cageo.2012.08.023>.
- Quinlan, J.R., 1986. Induction of decision trees. *Mach. Learn.* 1 (1), 81–106. <https://doi.org/10.1007/BF00116251>.
- Quinlan, J.R., 2003. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo.
- Reed, B.L., Menzie, W.D., McDermott, M., et al., 1989. Undiscovered lode tin resources of the Seward Peninsula, Alaska. *Econ. Geol.* 84 (7), 1936–1947. <https://doi.org/10.2113/gsecongeo.84.7.1936>.
- Robert, H.L., 1989. Theory of the backpropagation neural network. *Proceeding of the International Joint Conference on Neural Networks (IJCNN)* 1, 593–605.
- Rose, P.R., Everett, J.R., Merin, I.S., 1984. Possible basin centered gas accumulation, Roton Basin, Southern Colorado. *Oil Gas J.* 82 (10), 190–197.
- Saikia, P., Baruah, R.D., Singh, S.K., et al., 2020. Artificial neural Networks in the domain of reservoir characterization: A review from shallow to deep models. *Comput. Geosci.* 135, 104357. <https://doi.org/10.1016/j.cageo.2019.104357>.
- Sarailidis, G., Wagener, T., Pianosi, F., 2023. Integrating scientific knowledge into machine learning using interactive decision trees. *Comput. Geosci.* 170, 105248. <https://doi.org/10.1016/j.cageo.2022.105248>.
- Schölkopf, B., Smola, A., Müller, K.R., 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.* 10 (5), 1299–1319. <https://doi.org/10.1162/089976698300017467>.
- Shanmugam, G., 1997. The Bouma sequence and the turbidite mind set. *Earth Sci. Rev.* 42 (4), 201–229. [https://doi.org/10.1016/S0012-8252\(97\)81858-2](https://doi.org/10.1016/S0012-8252(97)81858-2).
- Singer, D.A., 1993. Basic concepts in three-part quantitative assessments of undiscovered mineral resources. *Nonrenewable Resour.* 2 (2), 69–81. <https://doi.org/10.1007/BF02272804>.
- Song, S., Mukerji, T., Hou, J., 2021. GANSim: Conditional facies simulation using an improved progressive growing of generative adversarial networks (GANs). *Math. Geosci.* 53 (7), 1413–1444. <https://doi.org/10.1007/s11004-021-09934-0>.
- Song, S., Mukerji, T., Hou, J., 2022. Bridging the gap between geophysics and geology with generative adversarial networks. *IEEE Trans. Geosci. Rem. Sens.* 60, 5902411. <https://doi.org/10.1109/TGRS.2021.3066975>.
- Sudre, C.H., Li, W., Vercauteren, T., et al., 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision*

- Support. Springer International Publishing, Cham, pp. 240–248. https://doi.org/10.1007/978-3-319-67558-9_28.
- Sun, C., Demyanov, V., Arnold, D., 2023. Geological realism in fluvial facies modelling with GAN under variable depositional conditions. *Comput. Geosci.* 27 (2), 203–221. <https://doi.org/10.1007/s10596-023-10190-w>.
- Sun, X., Opuencia, M.J.C., Alexandrovich, T.P., et al., 2022. Modeling and optimization of vegetable oil biodiesel production with heterogeneous nano catalytic process: Multi-layer perceptron, decision regression tree, and K-Nearest Neighbor methods. *Environ. Technol. Innov.* 27, 102794. <https://doi.org/10.1016/j.eti.2022.102794>.
- Szegedy, C., Liu, W., Jia, Y., et al., 2015. Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp. 1–9. <https://doi.org/10.1109/CVPR.2015.7298594>.
- Tissot, B.P., Welte, D.H., 1978. *Petroleum Formation and Occurrence*. Springer, Berlin. <https://doi.org/10.1007/978-3-642-96446-6>.
- Van Der Meer, F., Van Dijk, P., Van Der Werff, H., et al., 2002. Remote sensing and petroleum seepage: A review and case study. *Terra Nova* 14 (1), 1–17. <https://doi.org/10.1046/j.1365-3121.2002.00390.x>.
- Van Wees, J.D., Mijndieff, H., Lutgert, J., et al., 2008. A Bayesian belief network approach for assessing the impact of exploration prospect interdependency: An application to predict gas discoveries in the Netherlands. *AAPG Bull.* 92 (10), 1315–1336. <https://doi.org/10.1306/06040808067>.
- Vapnik, V., 2006. *Estimation of Dependences Based on Empirical Data*. Springer, New York, New York. <https://doi.org/10.1007/0-387-34239-7>.
- Vapnik, V., Golowich, S.E., Smola, A., 1996. Support vector method for function approximation, regression estimation and signal processing. In: *Proceedings of the 10th International Conference on Neural Information Processing Systems*. ACM, pp. 281–287.
- Virieux, J., Operto, S., 2009. An overview of full-waveform inversion in exploration geophysics. *Geophysics* 74 (6), WCC1–WCC26. <https://doi.org/10.1190/1.3238367>.
- Voordouw, G., 2011. Production-related petroleum microbiology: Progress and prospects. *Curr. Opin. Biotechnol.* 22 (3), 401–405. <https://doi.org/10.1016/j.copbio.2010.12.005>.
- Wahba, G., 1998. Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. In: *Advances in Kernel Methods*. The MIT Press. <https://doi.org/10.7551/mitpress/1130.003.0009>.
- Wainer, J., Cawley, G., 2021. Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Syst. Appl.* 182, 115222. <https://doi.org/10.1016/j.eswa.2021.115222>.
- Wang, F., Chen, D., Li, M., et al., 2024. A novel method for predicting shallow hydrocarbon accumulation based on source-fault-sand (S-F-Sd) evaluation and ensemble neural network (ENN). *Appl. Energy* 359, 122684. <https://doi.org/10.1016/j.apenergy.2024.122684>.
- Wang, J., Heng, J., Xiao, L., et al., 2017. Research and application of a combined model based on multi-objective optimization for multi-step ahead wind speed forecasting. *Energy* 125, 591–613. <https://doi.org/10.1016/j.energy.2017.02.150>.
- Wang, Q., Chen, D., Li, M., et al., 2023. A novel method for petroleum and natural gas resource potential evaluation and prediction by support vector machines (SVM). *Appl. Energy* 351, 121836. <https://doi.org/10.1016/j.apenergy.2023.121836>.
- Weiss, K., Khoshgoftaar, T.M., Wang, D., 2016. A survey of transfer learning. *J. Big Data* 3 (1), 9. <https://doi.org/10.1186/s40537-016-0043-6>.
- White, I.C., 1885. The geology of natural gas. *Sciences* 5 (125), 521–522. <https://doi.org/10.1126/science.ns-5.125.521>.
- Yu, H., Zhao, Z., Dahi Taleghani, A., et al., 2024. Modeling thermal-induced well-head growth through the lifecycle of a well. *Geoenergy Sci. Eng.* 241, 213098. <https://doi.org/10.1016/j.geoen.2024.213098>.
- Yue, D., Li, W., Wang, W., et al., 2019. Fused spectral-decomposition seismic attributes and forward seismic modelling to predict sand bodies in meandering fluvial reservoirs. *Mar. Petrol. Geol.* 99, 27–44. <https://doi.org/10.1016/j.marpetgeo.2018.09.031>.
- Zangqa, S., Saffou, E., Gholami, R., et al., 2024. Hydraulic fracturing potential of tight gas reservoirs: A case study from a gas field in the Bredasdorp Basin, South Africa. *Gas Sci. Eng.* 128, 205364. <https://doi.org/10.1016/j.jgsce.2024.205364>.
- Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: *Computer Vision – ECCV 2014*. Springer International Publishing, Cham, pp. 818–833. https://doi.org/10.1007/978-3-319-10590-1_53.
- Zhang, T.F., Tilke, P., Dupont, E., et al., 2019. Generating geologically realistic 3D reservoir facies models using deep learning of sedimentary architecture with generative adversarial networks. *Pet. Sci.* 16 (3), 541–549. <https://doi.org/10.1007/s12182-019-0328-4>.
- Zhang, Y., Liu, Y., Zhang, H., et al., 2020. Seismic facies analysis based on deep learning. *IEEE Geosci. Remote Sens. Lett.* 17 (7), 1119–1123. <https://doi.org/10.1109/LGRS.2019.2941166>.
- Zhao, T., 2018. Seismic facies classification using different deep convolutional neural networks. In: *SEG Technical Program Expanded Abstracts 2018*. Society of Exploration Geophysicists, pp. 2046–2050. <https://doi.org/10.1190/segam2018-2997085.1>.
- Zhao, Z., Chen, Y., Zhang, Y., et al., 2023. A deep learning model for predicting the production of coalbed methane considering time, space, and geological features. *Comput. Geosci.* 173, 105312. <https://doi.org/10.1016/j.cageo.2023.105312>.
- Zhen, X., Vinnem, J.E., Han, Y., et al., 2022. Development and prospects of major accident indicators in the offshore petroleum sector. *Process Saf. Environ. Prot.* 160, 551–562. <https://doi.org/10.1016/j.psep.2022.02.050>.
- Zhu, L., Bijeljic, B., Blunt, M.J., 2025. Diffusion model-based generation of three-dimensional multiphase pore-scale images. *Transp. Porous Medium.* 152 (3), 22. <https://doi.org/10.1007/s11242-025-02158-4>.