scientific reports

OPEN

Machine learning enhanced formation pressure prediction using integrated well logging and mud logging

Jiwen Liang[1], Ming Luo[1], Wentuo Li[1], Bo Sun[2]✉, Chuanliang Yan[2,3], Zhongying Han[2,3] & Yuanfang Cheng[2,3]

The difficulty of accurately predicting abnormally high-pressure formation pressure is one of the critical challenges in the field of petroleum engineering. Due to the low accuracy of formation pressure prediction and the narrow drilling safety density window, accidents such as leakage and blowout occur frequently. To address this issue, improving the accuracy of pore pressure predictions is essential. The well logging and mud logging data were combined to analyze the correlation between various parameters. Analysis using the Spearman correlation coefficient revealed that pore pressure exhibits varying correlation relationships with different parameters. Pore pressure is closely related to factors such as depth, weight of hanging, and mud weight. Pore pressure has a medium to high correlation with the rate of penetration, weight on bit, torque, slurry pump pressure, acoustic time difference, density, and volume of clay. Pore pressure has a medium to low correlation with the rotation per minute. Based on machine learning algorithms and a large amount of known data, a machine learning formation pressure model with integrated well logging and mud logging data (IWM) was established. The prediction results of traditional models and IWM models were compared using neighboring wells as the prediction targets. The results indicate that the backpropagation neural network model based on a genetic algorithm and IWM (IWM-GABP) achieves the highest prediction accuracy, with an average prediction accuracy greater than 96%. When predicting formation pressure, it is advisable to use the back propagation neural network model based on IWM or the IWM-GABP model, rather than the radial basis function neural network model based on IWM. The IWM model significantly reduces the prediction error of formation pore pressure, achieving an average improvement of 8.32% enhancement in prediction accuracy compared to traditional data models. The research method effectively improves the accuracy of formation pressure prediction and provides support for efficient on-site development.

Keywords Pore pressure, Accurate prediction, Integrated logging data, Abnormally high-pressure, Machine learning

Abbreviations

ADT	Acoustic time difference
BP	Back propagation
CNN	Convolutional neural network
GABP	Back propagation model with genetic algorithm
IWM-BP	Integrated data model based on the BP neural network
IWM-CNN	Integrated data prediction model based on the CNN
IWM-GABP	Integrated data model based on the GABP neural network
IWM-RBF	Integrated data prediction model based on the RBF
IWM-RF	Integrated data prediction model based on the RF
IWM-SVM	Integrated data prediction model based on the SVM
MW	Mud weight
Pp	Pore pressure

[1]Hainan Branch of CNOOC (China) Co., Ltd, Haikou 570100, Hainan, China. [2]School of Petroleum Engineering, China University of Petroleum (East China), Qingdao 266580, Shandong, China. [3]State Key Laboratory of Deep Oil and Gas, China University of Petroleum (East China), Qingdao 266580, China. ✉email: sunupcb@163.com

RBF	Radial basis function
RF	Random forest algorithm
ROP	Rate of penetration
RPM	Rotation per minute
SPP	Slurry pump pressure
SVM	Support vector machine
Tor	Torque
VCL	Volume of clay
WOB	Weight on bit
WOH	Weight of hanging

The problem of accurate prediction of formation pressure has always been one of the focuses in the field of petroleum engineering[1–5], especially in areas with abnormally high pressure. According to statistical data, abnormally high-pressure accounts for about 48% of oil and gas fields worldwide[6]. The parameters of formation pore pressure are the basis for drilling and completion engineering and oil and gas field development. Formation pore pressure is related to safe, fast, effective, and economical drilling and completion[7–9], and even determines the success or failure of drilling. Risks often occur due to the low prediction accuracy of abnormally high-pressure wells, mainly including blowout, wellbore instability, wellbore scrapping, casing waste, and a surge in drilling economic costs[10–13]. Accurate prediction of formation pressure/pore pressure has become crucial to address many oil and gas production related problems[14–16]. When studying the formation pressure of high-temperature and high-pressure wells in the South China Sea, traditional methods have low prediction accuracy and cannot meet the requirements of safe drilling. When exploring solutions, practical research has discovered a simple and significantly improved method for predicting formation pressure, which is the machine learning prediction method based on integrated data.

The study of pore pressure prediction is a complex and challenging task. In recent years, experts and scholars at home and abroad have conducted a large amount of research on pore pressure prediction methods. However, due to the high development of computers, the current prediction methods mainly include two categories.

The first type of prediction is based on empirical or semi-empirical formulas, which are generally fitted and solved based on on-site well logging or mud logging data. The main representatives include the Eaton method, which establishes normal trend lines based on well logging acoustic data[17–20], and the Bowers method, which distinguishes and predicts pressure mechanisms based on well logging acoustic and density data[21–25]. The Fan method[26,27] for calculating the correlation between well logging acoustic waves, porosity, and effective stress, and the Dc index method[28–30] for calculating formulas based on partial mud logging parameters and data. However, the parameter selection of these empirical or semi-empirical methods is subjective and often leads to significant prediction errors, resulting in poor on-site application results.

The second type is prediction methods mainly based on intelligent algorithms, including neural networks, random forests, deep learning, etc.[31–35]. Currently, there is relatively little research on these methods, and most of them focus on using logging data for pore pressure prediction, without fully exploring and utilizing favorable on-site data (including well logging and mud logging data). While existing research has investigated the combination of mud logging and well logging data[36], the study has limitations such as small sample sizes, shallow formation depth, inadequate predictive evaluation, and the applicability of its conclusions may be restricted by geographical factors. Therefore, there are still many shortcomings in the second type of research, and their methods often have poor results in predicting formation pressure.

Therefore, to improve the accuracy of predicting formation pore pressure and make up for the shortcomings of current research. Well logging and mud logging data are combined in this study to analyze the correlation between various parameters. Based on machine learning algorithms and a large amount of known data, a machine learning formation pressure model with integrated well logging and mud logging data (IWM) was established. The model is predicted and applied to neighboring wells with brand-new data. At the same time, to evaluate the computational accuracy and advantages of the model, conventional machine learning well logging or mud logging models were compared, and the calculation results and accuracy of various data models were obtained. Comprehensively evaluated the integrated data model established in this study. The research method can effectively improve the accuracy of reservoir pressure prediction and provide support for efficient on-site development.

Methods and data handling

The establishment of IWM datasets and prediction models is the foundation for accurate machine learning prediction of formation pressure. Excellent machine learning model calculation accuracy is often based on accurate and rich datasets. This study established a machine learning prediction model for formation pressure based on well logging and mud logging data using an integrated dataset. The model can then be used for formation pressure prediction and accuracy evaluation and optimization.

Therefore, a prediction dataset and a modeling dataset were established based on well logging and mud logging data. The study conducted data correlation analysis and normalization on the modeling dataset and processed the formatted data. Subsequently, the processed data is randomized and sorted, and training and validation sets are established proportionally. Different machine learning models are selected for training. When the training accuracy cutoff condition is met, the training is completed and the optimal model and random order relationship are obtained. Finally, the model and its relationship are used to perform weighted prediction analysis and inverse normalization on the predicted data. The modeling and processing flow of this study is shown in Fig. 1.

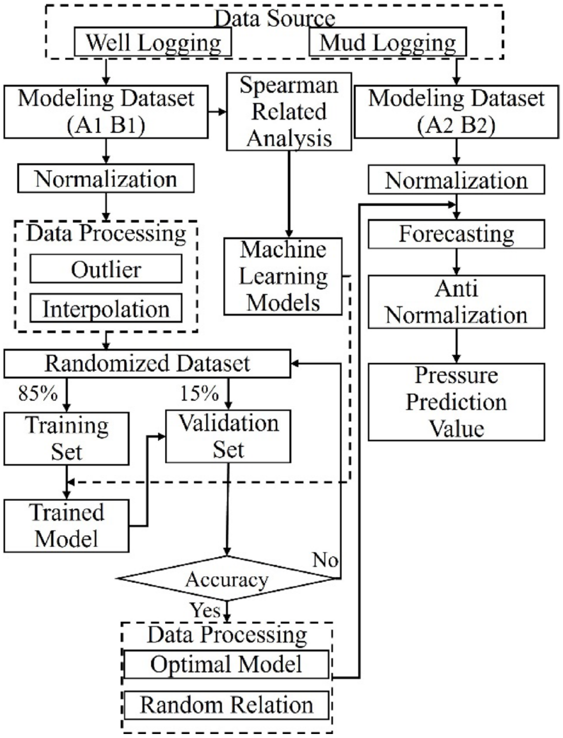


Fig. 1. The process of research and analysis.

Data source and relationship analysis

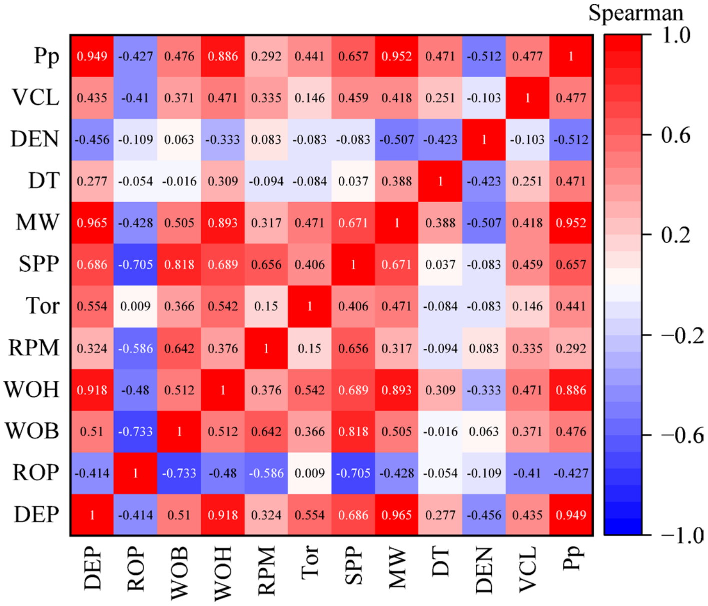
It is often considered reasonable to predict unknown adjacent wells by utilizing modeling and fitting information from known drilling data. This approach helps avoid overfitting of data and enhances the reliability or credibility of evaluation studies. All data used in this study originates from the Ledong and Dongfang blocks in the Yingqiong Basin, China. A notable feature of this region is the presence of strata with abnormally high pore pressure. The known well dataset used for modeling in the Ledong block is designated as A1, while the known well dataset used for modeling in the Dongfang block is designated as B1. The adjacent well datasets used for final prediction and evaluation are designated as A2 and B2, respectively.

The IWM dataset and prediction accuracy evaluation validation dataset mainly include well logging and mud logging data. The key parameters of logging data mainly include vertical depth (DEP), rate of penetration (ROP), weight on bit (WOB), weight of hanging (WOH), rotation per minute (RPM), torque (Tor), slurry pump pressure (SPP), mud weight (MW), etc. The key parameters of logging data mainly include acoustic time difference (ADT), density (DEN), and volume of clay (VCL) Wait. The integrated dataset used for model training, validation, and prediction is a data matrix composed of mud logging parameters and well logging parameters. The first eleven columns of the matrix are vertical depth, rate of penetration, weight on bit, weight of hanging, rotation per minute, torque, slurry pump pressure, mud weight, acoustic time difference, density, and volume of clay respectively. The last column of the matrix contains pore pressure data. The input condition of the model is the first eleven columns of the data matrix. The output result is pore pressure (Pp). Therefore, the established data input and output matrix relationship is shown in Eq. (1).

$$Pp = f([DEP\ ROP\ WOB\ WOH\ RPM\ Tor\ SPP\ MW\ ADT\ DEN\ VCL]) \tag{1}$$

At the same time, to select the correlation characteristics of the parameters, correlation analysis was conducted on each parameter in the dataset, and a correlation coefficient matrix was established. The main methods for analyzing parameter correlation include Pearson, Spearman, and Kendall methods. The Pearson method generally requires parameters to satisfy continuity and normal distribution characteristics, and variables are linearly correlated, so the requirements for use are relatively strict; The Spearman method and Kendall method generally do not consider the sample distribution morphology and are general nonparametric methods. In this study, the Spearman method was used for correlation analysis, and the calculation formula is shown in Eq. (2). The correlation coefficient matrix obtained through calculation is shown in Fig. 2.

$$\rho_j = \frac{\sum_{i=1}^{i=n} (x_{ij} - \overline{x}_j)\left(p_{i_j} - \overline{p}_j\right)}{\sqrt{\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2 \sum_{i=1}^{n} \left(p_{ij} - \overline{p}_j\right)^2}} \tag{2}$$



	DEP	ROP	WOB	WOH	RPM	Tor	SPP	MW	DT	DEN	VCL	Pp
Pp	0.949	-0.427	0.476	0.886	0.292	0.441	0.657	0.952	0.471	-0.512	0.477	1
VCL	0.435	-0.41	0.371	0.471	0.335	0.146	0.459	0.418	0.251	-0.103	1	0.477
DEN	-0.456	-0.109	0.063	-0.333	0.083	-0.083	-0.083	-0.507	-0.423	1	-0.103	-0.512
DT	0.277	-0.054	-0.016	0.309	-0.094	-0.084	0.037	0.388	1	-0.423	0.251	0.471
MW	0.965	-0.428	0.505	0.893	0.317	0.471	0.671	1	0.388	-0.507	0.418	0.952
SPP	0.686	-0.705	0.818	0.689	0.656	0.406	1	0.671	0.037	-0.083	0.459	0.657
Tor	0.554	0.009	0.366	0.542	0.15	1	0.406	0.471	-0.084	-0.083	0.146	0.441
RPM	0.324	-0.586	0.642	0.376	1	0.15	0.656	0.317	-0.094	0.083	0.335	0.292
WOH	0.918	-0.48	0.512	1	0.376	0.542	0.689	0.893	0.309	-0.333	0.471	0.886
WOB	0.51	-0.733	1	0.512	0.642	0.366	0.818	0.505	-0.016	0.063	0.371	0.476
ROP	-0.414	1	-0.733	-0.48	-0.586	0.009	-0.705	-0.428	-0.054	-0.109	-0.41	-0.427
DEP	1	-0.414	0.51	0.918	0.324	0.554	0.686	0.965	0.277	-0.456	0.435	0.949

Fig. 2. Spearman correlation coefficient matrix.

where ρ_j is the combination relationship, x_{ij} is the variable value of the combination relationship, $\overline{x}_j$ is the mean value of the combination relationship variable, p_{ij} is the observation value of the combination relationship, and $\overline{p}_j$ is the mean of the combination relationship observation value, $j = 1, 2, 3, \ldots 11, 12$.

According to the Spearman correlation coefficient matrix, it can be found that pore pressure has different correlation relationships with various parameters. Pore pressure has a strong correlation with DEP, WOH, and MW. While it has a medium to high correlation with ROP, WOB, TOR, SPP, ADT, DEN, and VCL. Pore pressure has a medium to low correlation with RPM.

Data processing and partitioning

The data used for model training in this study consists of 6520 entries, each containing 12 parameters as shown in Eq. (1). The initial dataset for training the model comprises a total of 78,240 data points. The statistical parameters of data can provide a comprehensive overview of the data, which is crucial for the transparency and credibility of research. Therefore, to clearly describe data characteristics and evaluate data quality. This study provides some statistical parameters of the research data, mainly including mean value, standard deviation, median value, and quartiles. The statistical results are shown in Supporting Table 1.

The unified processing of data mainly includes removing outliers, completing missing values, randomizing data, partitioning the dataset, and normalizing the data.

Firstly, to make the training model more reliable, the outliers in the dataset are treated as 3σ[37]. Abnormal data outside the range of three standard deviations were excluded by using this principle. The data on standard deviation can be found in Supporting Table 1. 39 and 26 abnormal data were removed from the A1 and B1 datasets, respectively. At the same time, for partially missing data, interpolation of neighboring values is used to complete.

Secondly, this study trained and validated the pore pressure prediction model using partitioned data from wells A1 and B1. The accuracy and rationality of model training are crucial. Randomly dividing and selecting the training set has many benefits[38]. It reduces the bias caused by data order or specific patterns and makes the model more generalizable. It helps to avoid overfitting the model to specific training data patterns. Meanwhile, randomly selecting partitioned datasets is a cross-validation method that can help evaluate the robustness of models on different datasets. Therefore, 85% of the data from datasets A1 and B1 were randomly selected as the model training data to train the machine learning model. The remaining data (15%) from datasets A1 and B1 were selected as the model validation test data to validate the model. Among them, Well A1 has a total of 4046 pieces of data. 3439 pieces of data were randomly selected for model training, and the remaining 607 pieces of data were selected for model validation; B1 well has a total of 2409 data pieces, of which 2047 were randomly selected for model training, and the remaining 362 were selected for model validation. At the same time, establish datasets A2 and B2 as input conditions for evaluating the accuracy of the trained prediction model, to evaluate the reliability or accuracy of the model. Among them, the A2 well has 698 data pieces, and the B2 well has 1918 data pieces.

Finally, to eliminate the influence of parameter magnitude on the model results and obtain reasonable input feature parameter values, all dataset parameters are normalized. After the model is established, reverse normalization can be performed to obtain the parameter results of the original morphological features.

Modeling and prediction with integrated well logging and mud logging data

To evaluate the accuracy and model optimization of the IWM model, this study established commonly used machine learning models based on IWM, including the back propagation (BP) neural network, the support vector machine (SVM), the BP model with genetic algorithm (GABP), the random forest algorithm (RF), the radial basis function (RBF) neural network, and the convolutional neural network (CNN). The IWM model was trained, validated, and used to predict formation pressure through recorded data. The following machine learning models are used to train and verify the existing data to achieve the best and approximate training and verification effect. Based on the approximate training effect, another prediction data set is applied. The accuracy of each model is evaluated by observing the prediction results.

Modeling with integrated well logging and mud logging data

Construction of IWM prediction model based on the BP

The neural network algorithm is one of the algorithms with self-learning ability, high-speed search for optimal solutions, and good practical application effects. The Back Propagation Neural Network was proposed by Rumelhart et al.[39], which has excellent function approximation ability through iteration and updating. According to the BP model structure, an integrated data model based on the BP neural network (IWM-BP) was established, as shown in Fig. 3. The main steps involved in the implementation are as follows: random initialization of weights, forward propagation calculation, loss calculation, backpropagation, and weight updates. Both the forward and reverse modules have undergone weight training. The loss function is trained using synthetic data. A crucial aspect of training the weights is the gradient calculated during backpropagation. These gradients are utilized to adjust the weights in each iteration, gradually reducing the loss function. In this study, the integrated data model for measurement and recording has 11 input neurons, hidden layer ranges of 2, 3, and 5 layers, one output neuron, a maximum iteration of 2000, and a learning rate of 0.01.

Construction of IWM prediction model based on the GABP

The genetic algorithm theory is an algorithm designed based on Darwin's theory of evolution, which can cause data to evolve in a positive direction to obtain the optimal solution[40]. This study established an integrated data model based on the GABP neural network (IWM-GABP). The input data was normalized measurement and recording integrated data. Firstly, the BP neural network and genetic algorithm parameters (including genetic algebra, race size, etc.) were initialized, and then fitness and iterative calculations were repeated to determine the initial weights and thresholds of the optimized BP network. Finally, the weights and thresholds were input into the BP model for training and validation, The model established through this process can be used for predicting formation pressure. This model obtains a more adaptive model by changing the population size and genetic algebraic parameters and uses this to predict pore pressure.

Construction of IWM prediction model based on the SVM

The support vector machine is a classic binary model[41], which works by identifying decision hyperplanes and completing data planning. A more common model is the linear SVM model. This study uses a nonlinear SVM model to handle input nonlinear problems. The nonlinear problem is transformed into a linear problem in the feature space through a Gaussian kernel function, and a linear SVM model is used in a high-dimensional feature space. Therefore, the decision function of the nonlinear SVM model is obtained as shown in Eq. (3). SVM itself does not have a backpropagation mechanism. This study obtained a more adaptive model by changing the penalty factor and radial basis function parameters and established an integrated data prediction model based on the SVM (IWM-SVM) to predict pore pressure.

$$f(x) = sign\left(\sum_{i=1}^{n} \alpha_i^* y_i \exp\left(-\left\| x - \boldsymbol{x}_i \right\|^2 / 2\sigma^2\right) + b^*\right) \tag{3}$$

where α_i^* is the optimal solution of the Lagrange multiplier, y_i is the class marker, x_i is the eigenvector, σ is the width parameter of the Gaussian function, and b^* is the optimal intercept.

Construction of IWM prediction model based on the RF

The random forest algorithm is one of the most used and powerful supervised learning algorithms[42], which can obtain regression prediction problems through the output of decision trees. This algorithm has obvious advantages, high accuracy, and processing large amounts of data without the need for dimensionality reduction design. The input condition of the model is standardized IWM, and a more adaptive model is obtained by

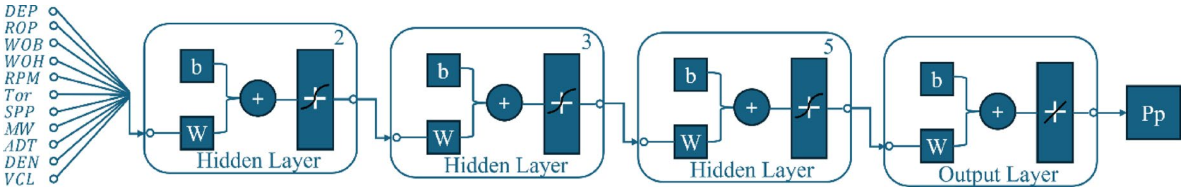


Fig. 3. IWM-BP model structure.

changing the number of decision trees and leaves parameters, which is used for predicting pore pressure. The algorithm does not have a display weight and backpropagation mechanism. The construction of trees is based on random samples and feature selection. This study obtained a more adaptive model by changing the decision tree parameters, and an integrated data prediction model based on the RF (IWM-RF) was established to predict pore pressure.

Construction of IWM prediction model based on the RBF
The radial basis function neural network is a structure activated by radial basis functions, typically consisting of input layer, hidden layer, and output layer[43]. The input layer of the model in this study is the standardized measurement and recording IWM matrix, the transformation functions of each unit in the hidden layer are radial basis functions, and the activation function in this study is the Gaussian function, as shown in Eq. (4). The function of the hidden layer is to map the data input to a high-dimensional space and then perform fitting. The function of the output layer is to weight the data calculated by the hidden layer. The weights of the output layer are adjusted through the gradient descent method. Finally, the RBF neural network structure was obtained, as shown in Eq. (5). The structure of the integrated data prediction model based on the RBF (IWM-RBF) in this study is shown in Fig. 4.

$$R\left(x_j, x_i\right) = \exp\left(-\frac{1}{2\sigma^2}\left\|x_{\ j} - x_i\right\|^2\right) \tag{4}$$

$$F(x) = \sum_{i=1}^{n} w_{ij} \exp\left(-\frac{1}{2\sigma^2}\left\|x_{\ j} - x_i\right\|^2\right) \tag{5}$$

where x_j is the input sample, x_i is the center point, $\left\|x_{\ j} - x_i\right\|$ is the norm, and w_{ij} is the weighted value.

Construction of IWM prediction model based on the CNN
The convolutional neural network is a deep learning algorithm that generally includes operations such as convolution, pooling, activation, loss, and output[44]. This study is based on CNN and integrated measurement and recording data. The data is tiled into 11 × 1 × 1 images and zero-centered, and then subjected to two-dimensional convolution to form 16 feature maps with a convolution kernel size of 5 × 1. Then, an optimized parametric rectified linear unit is used as the activation function to perform nonlinear mapping on the data, which avoids the disadvantage of a gradient of zero in the negative range, as shown in Eq. (6). Then, perform a second convolution to form 32 feature maps and activate them. Form a fully connected layer and a regression layer for model training, set the maximum number of model training times to 800, and reduce the learning rate by half after 400 times, with an initial learning rate of 0.01. The model training utilizes a gradient descent algorithm to update the weights of convolutional kernels and fully connected layers, and an integrated data prediction model based on the CNN (IWM-CNN) was established.

$$f(x) = \max(ax, x) \tag{6}$$

where *a* is the backpropagation learning parameter.

Model training and validation

The hyperparameters for training the model are accuracy values, as shown in Fig. 1. Hyperparameters start from 100% during the training process. If the hyperparameter requirements are still not met after reaching the iteration number, the hyperparameter value will be reduced and the next iteration calculation will be carried out. The training will be stopped when the optimal parameter values are obtained. The hyperparameter reduction step size is 0.01%. Therefore, the accuracy value of the training model obtained when the training stops is the hyperparameter value, also known as the training set goodness of fit in the study.

The study established an IWM-BP model through the computer and optimized model training on the A1 and B1 well datasets. Finally, the model training process and results were obtained.

According to the model, after data training for well A1, the goodness of fit of the training set was 0.9956, and the goodness of fit of the validation test set was 0.9941. The best validation performance is 0.0032146, achieved at epoch 17. The regression state and validation performance of the training process data is shown in Fig. 5, and the results of the model training data and validation data are shown in Fig. 6. The results of the B1 well

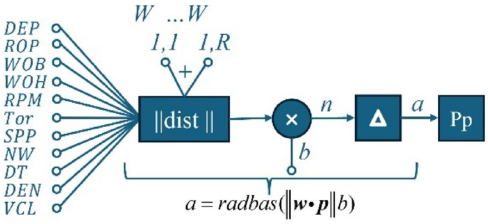
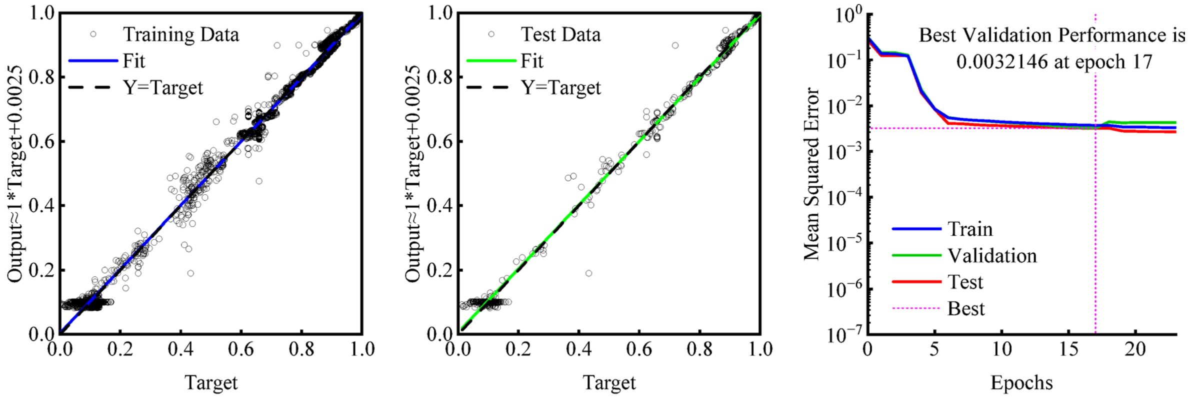
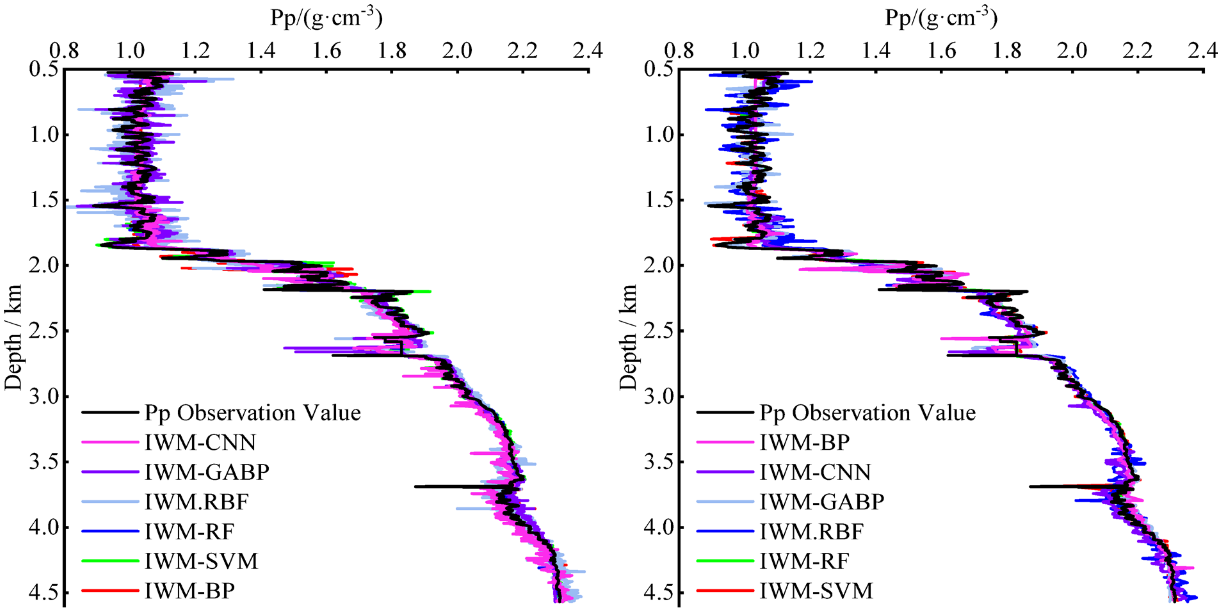


Fig. 4. IWM-RBF model structure.



(a)Training data (b)Test data (c)Best validation performance

Fig. 5. Regression state and validation performance.



(a)Training data (b)Test data

Fig. 6. A1 well model training data and validation data results.

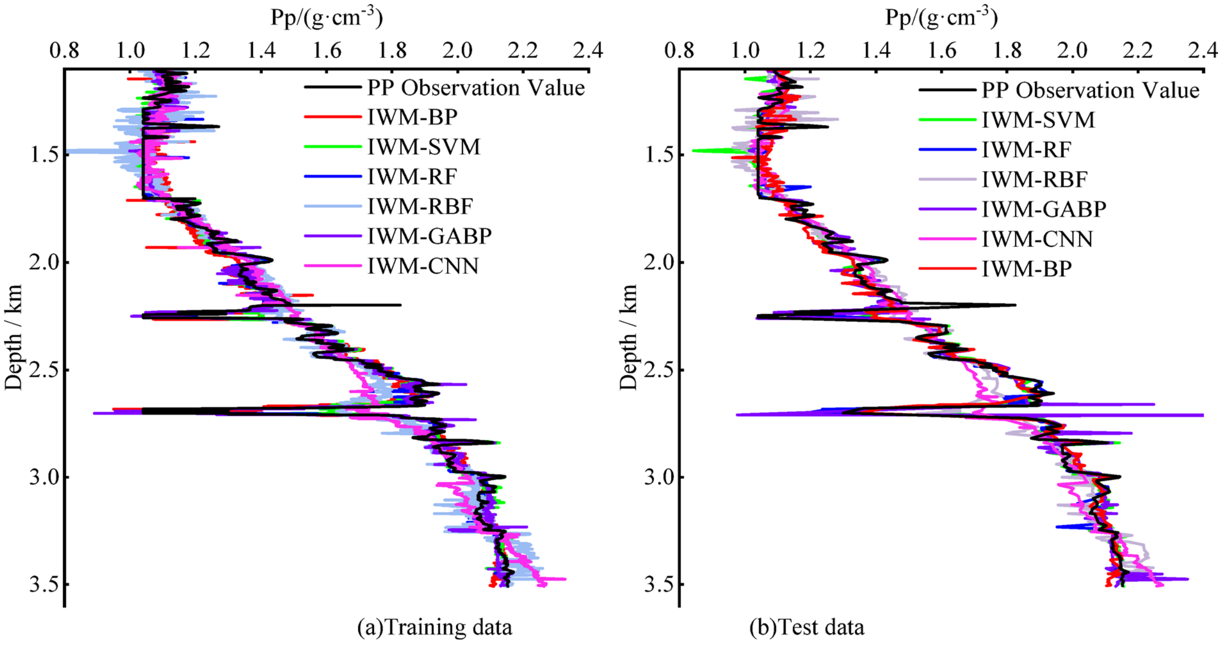
model training and validation data are shown in Fig. 7. The goodness of fit of the training set is 0.9740, and the goodness of fit of the validation test set is 0.9763. According to the observed values of goodness of fit, the model training results are in line with expectations.

Similarly, data based on IWM and other models (GABP, SVM, RF, RBF, CNN) were established separately. The optimal model training was performed on the A1 and B1 well datasets to obtain the model training process and results. The training and validation data of the A1 and B1 well models are shown in Figs. 6 and 7. After calculation, the goodness of fit of the A1 well training set was 0.9958, 0.9988, 0.9991, 0.9911, and 0.9922, respectively. The goodness of fit of the A1 well testing validation set was 0.9952, 0.9986, 0.9987, 0.9908, and 0.9927, respectively. The goodness of fit of the B1 well training set were 0.9909, 0.9672, 0.9943, 0.9496, and 0.9391, respectively. The goodness of fit of the B1 well testing validation set were 0.9093, 0.9768, 0.9908, 0.9357, and 0.9411, respectively. Based on the observed values of goodness of fit for each result, the model training results also meet expectations.

Prediction of formation pore pressure

Based on the established and trained IWM models, pore pressure prediction was performed on adjacent wells A2 and B2. The prediction results are shown in Fig. 8. Based on the calculation results, the prediction accuracy of each model was calculated, as shown in Table 1.

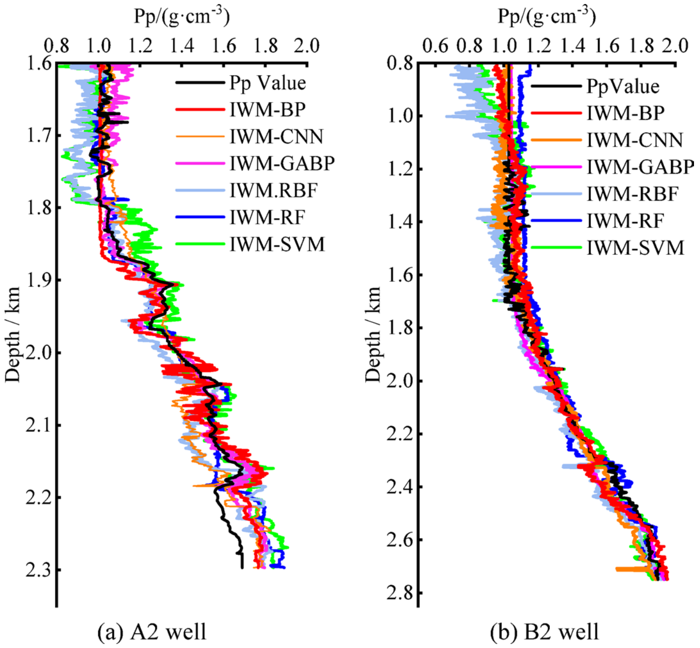
According to Fig. 8 of the model prediction results and Table 1 of the prediction accuracy calculation, it can be found that except for IWM-RBF, the prediction accuracy of all models is greater than 90%. The IWM-



(a)Training data

(b)Test data

Fig. 7. B1 well model training data and validation data results.



(a) A2 well

(b) B2 well

Fig. 8. Prediction results of pore pressure.

Model	A2 well prediction accuracy (%)	B2 well prediction accuracy (%)
IWM-BP	92.08	94.91
IWM-GABP	94.66	97.98
IWM-SVM	93.80	94.89
IWM-RF	90.52	95.14
IWM-RBF	85.16	89.52
IWM-CNN	90.32	96.31

Table 1. Prediction accuracy of each model.

GABP model has the highest prediction accuracy. The average prediction accuracy is over 96%. When predicting formation pressure, it is advisable to use IWM-BP or IWM-GABP models, and it is not advisable to use IWM-RBF models.

Comparison and evaluation of prediction effects

Simply looking at the prediction results and accuracy of formation pressure made above cannot intuitively demonstrate the advantages of the integrated data model in this study. Therefore, to evaluate the advantages and calculation accuracy of the IWM data model, this study compared conventional machine learning logging and logging models and obtained the calculation results of various data models.

The calculation results and comparison of the data models for wells A2 and B2 are shown in Figs. 9 and 10. Based on the calculation results, the prediction accuracy of each model was calculated, as shown in Fig. 11.

According to the calculation results, for well A2, it can be found that the IWM-GABP model has the highest prediction accuracy, with a formation pressure prediction error of 5.34%, the IWM-SVM model has the lowest calculation accuracy, and the formation pressure prediction error is 16.20%. Compared to traditional logging and logging data models, the calculation errors of the IWM-BP model are reduced by 6.21% and 5.82%, respectively. The calculation errors of the IWM-SVM model are reduced by 41.7% and 11.77%, respectively. The calculation errors of the IWM-GABP model are reduced by 1.96% and 4.52%, respectively. The calculation errors of the IWM-RF model are reduced by 3.50% and 7.48%, respectively. The calculation errors of the IWM-RBF model are reduced by 19.93% and 16.97%, respectively. The calculation errors of the IWM-CNN model are reduced by 3.50% and 1.74%, respectively. Overall, compared to traditional methods, the prediction accuracy of this model has improved by an average of 12.80% and 8.05%.

According to the calculation results, for well B2, it can be found that the IWM-GABP model has the highest prediction accuracy, with a formation pressure prediction error of 2.02%, the IWM-RBF model has the lowest calculation accuracy, and the formation pressure prediction error is 10.47%. Compared to traditional logging and logging data models, the calculation errors of the IWM-BP model are reduced by 11.32% and 9.58%, respectively. The calculation errors of the IWM-SVM model are reduced by 2.81% and 10.28%, respectively. The calculation errors of the IWM-GABP model are reduced by 4.49% and 6.47%, respectively. The calculation errors of the IWM-RF model are reduced by 1.25% and 3.17%, respectively. The calculation errors of the IWM-RBF model are reduced by 3.96% and 12.51%, respectively. The calculation errors of the IWM-CNN model are

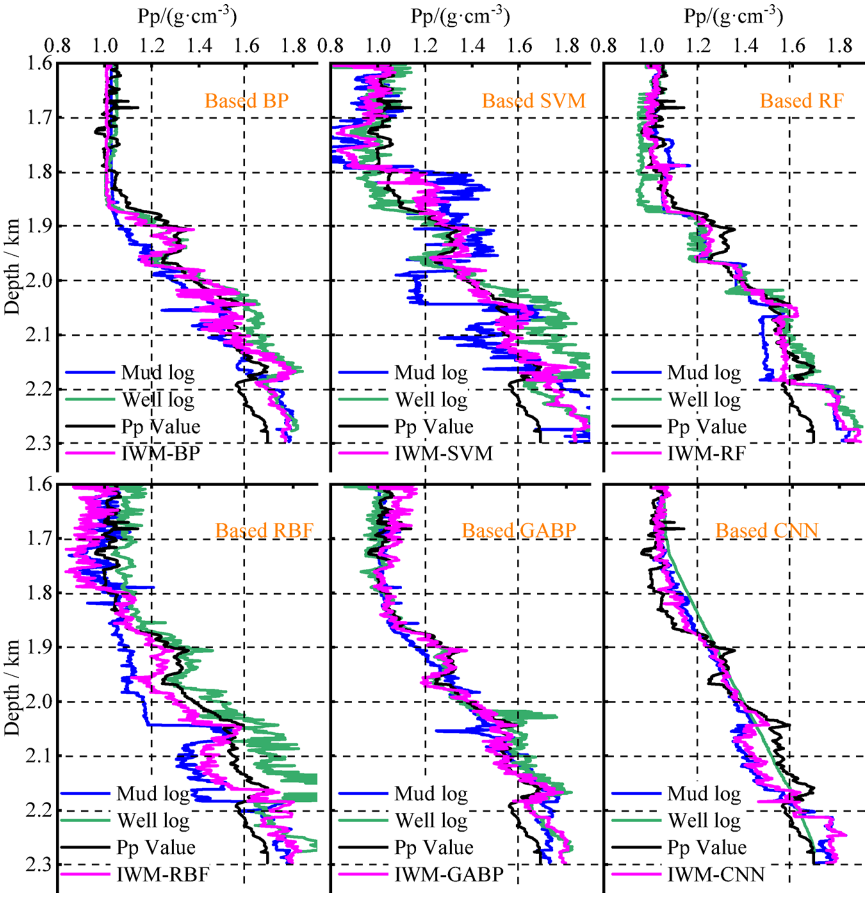


Fig. 9. Comparison of prediction results for well A2.

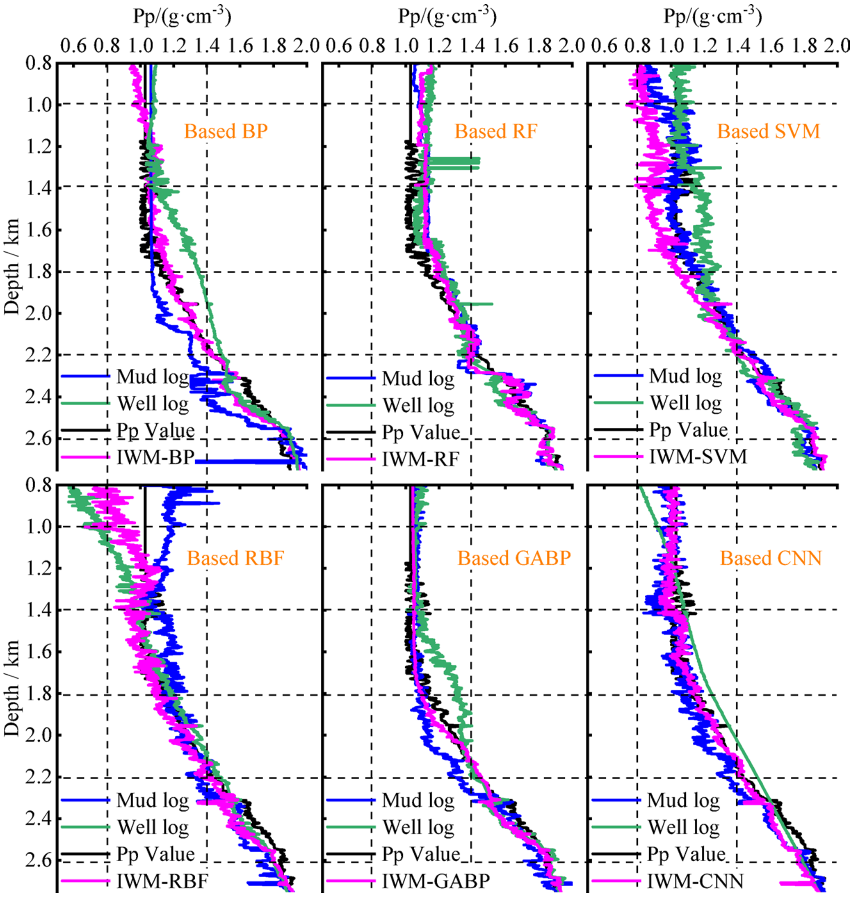


Fig. 10. Comparison of prediction results for well B2.

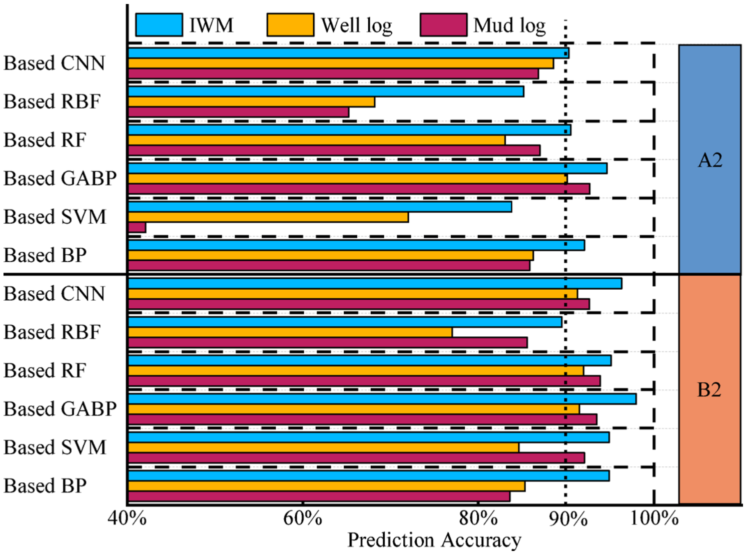


Fig. 11. Comparison bar chart of prediction accuracy.

reduced by 3.64% and 5.01%, respectively. Overall, compared to traditional methods, the prediction accuracy of this model has improved by an average of 4.58% and 7.84%.

Therefore, it can be found that compared to traditional mud logging or well logging data models, the integrated mud logging and well logging data model has higher accuracy in predicting formation pressure, with an average improvement of 8.69% and 7.95% in predicting formation pore pressure.

Conclusions and discussion

The study integrates mud logging and well logging data to analyze relationships among various parameters. Based on commonly used machine learning models for formation pressure, a formation pressure prediction model was established based on integrated combination data. Formation pressure prediction and analysis were conducted, and the accuracy of the integrated data model was evaluated by comparing it with traditional data methods. Research has found that:

1. Analysis using the Spearman correlation coefficient revealed that pore pressure exhibits varying correlation relationships with different parameters. Pore pressure is closely related to factors such as depth, weight of hanging, and mud weight. Pore pressure has a medium to high correlation with the rate of penetration, weight on bit, torque, slurry pump pressure, acoustic time difference, density, and volume of clay. Pore pressure has a medium to low correlation with the rotation per minute.
2. All integrated data models have a prediction accuracy exceeding 90%. The integrated data model utilizing the back propagation neural network method combined with a genetic algorithm achieves the highest prediction accuracy, averaging over 96% accuracy. To predict formation pressure, it is advisable to use an ensemble data model employing the backpropagation neural network method, ideally in combination with a genetic algorithm. Using the integrated data prediction model based on the radial basis function method is not advisable.
3. A comparative evaluation and analysis revealed that the integrated combination data model outperforms traditional mud logging and well logging data models. It not only predicts the formation pressure but also maintains high prediction accuracy. Specifically, there is an average improvement of 8.32% in predicting formation pressure accuracy.

According to the results, this model indeed has high prediction accuracy. However, this model also has certain limitations in field applications. Current machine learning models' parameters or algorithm layers may not apply to all blocks worldwide, but the methods proposed in current research are universally applicable. Different intelligent models should be trained in different regions. In practical applications, some older wells lack advanced measurement techniques. There may be a situation of insufficient parameters, which require specific analysis based on the situation. In addition, adding new data parameters may also improve prediction accuracy, as long as the parameter is meaningful and economically measurable. However, overall, this method is meaningful for improving the accuracy of reservoir pressure prediction and provides new ideas. Of course, if this method can be improved in the future to establish a real-time prediction model for formation pressure, it may be more suitable for engineering practice.

Data availability

The datasets used and analyzed during the current study available from the corresponding author on reasonable requset.

Received: 31 July 2024; Accepted: 12 June 2025
Published online: 02 July 2025

References

1. Zhang, G. et al. A robust approach to pore pressure prediction applying petrophysical log data aided by machine learning techniques. *J. Energy Rep.* **8**, 2233–2247. https://doi.org/10.1016/j.egyr.2022.01.012 (2022).
2. Li, C. et al. New understanding of overpressure responses and pore pressure prediction: Insights from the effect of clay mineral transformations on mudstone compaction. *J. Eng. Geol.* **297**, 106493. https://doi.org/10.1016/j.enggeo.2021.106493 (2022).
3. Shi, M. et al. Review and prospect prediction technology for formation pore pressure by geophysical well logging. *J. Prog. Geophys.* **35**, 1845–1853. https://doi.org/10.6038/pg2020DD0435 (2020).
4. Baouche, R. et al. Assessment of reservoir stress state and its implications for Paleozoic tight oil reservoir development in the Oued Mya Basin, northeastern Algerian Sahara. *J. Geosyst. Geoenviron.* **2**, 100112. https://doi.org/10.1016/j.geogeo.2022.100112 (2023).
5. Baouche, R. et al. Petrophysical, geomechanical and depositional environment characterization of the Triassic TAGI reservoir from the Hassi Berkine South field, Berkine Basin, Southeastern Algeria. *J. Nat. Gas Sci. Eng.* **92**, 104002. https://doi.org/10.1016/j.jngse.2021.104002 (2021).
6. Wang, X., Feng, B. & Li, X. Application of prediction by well-constrained seismic pressure in rolling exploration and development. *J. Oil Geophys. Prospect.* **37**, 391–394 (2002).
7. Xie, Y. & Gao, Y. Recent domestic exploration progress and direction of CNOOC. *J. China Pet. Explor.* **25**, 20–30 (2020).
8. Zhang, J. Pore pressure prediction from well logs: Methods, modifications, and new approaches. *J. Earth Sci. Rev.* **108**, 50–63. https://doi.org/10.1016/j.earscirev.2011.06.001 (2011).
9. Lockhart, L. P., Flemings, P. B., Nikolinakou, M. & Germaine, J. Velocity-based pore pressure prediction in a basin with late-stage erosion: Delaware Basin, US. *J. Mar. Pet. Geol.* **150**, 106159. https://doi.org/10.1016/j.marpetgeo.2023.106159 (2023).
10. Baouche, R. et al. Characterization of pore pressure, fracture pressure, shear failure and its implications for drilling, wellbore stability and completion design—A case study from the Takouazet field, Illizi Basin, Algeria. *J. Mar. Pet. Geol.* **120**, 104510. https://doi.org/10.1016/j.marpetgeo.2020.104510 (2020).
11. Li, Z., Xie, R. & Yuan, J. Study on the drilling safety probability interval in narrow pressure window formation in deepwater HPHT gas fields. *J. Nat. Gas Ind.* **40**, 88–93. https://doi.org/10.3787/j.issn.1000-0976.2020.12.010 (2020).
12. Lian, T., Fan, H. & Yu, L. Application of logging constrained seismic inversion in predicting formation pore pressure. *J. West China Explor. Eng.* **21**, 64–65. https://doi.org/10.3969/j.issn.1004-5716.2009.06.022 (2009).
13. Ezeakacha, C. P., Salehi, S. & Kiran, R. Lost circulation and filter cake evolution: Impact of dynamic wellbore conditions and wellbore strengthening implications. *J. Pet. Sci. Eng.g.* **171**, 1326–1337. https://doi.org/10.1016/j.petrol.2018.08.063 (2018).
14. Ganguli, S. S. et al. Deep thermal regime, temperature induced over-pressured zone and implications for hydrocarbon potential in the Ankleshwar oil field, Cambay basin, India. *J. Asian Earth Sci.* **161**, 93–102. https://doi.org/10.1016/j.jseaes.2018.05.005 (2018).
15. Mahetaji, M. & Brahma, J. Prediction of minimum mud weight for prevention of breakout using new 3D failure criterion to maintain wellbore stability. *J. Rock Mech. Rock Eng.* **57**, 2231–2252. https://doi.org/10.1007/s00603-023-03679-4 (2024).

16. Ganguli, S. S. & Sen, S. Investigation of present-day in-situ stresses and pore pressure in the south Cambay Basin, western India: Implications for drilling, reservoir development and fault reactivation. *J. Mar. Pet. Geol.* **118**, 104422. https://doi.org/10.1016/j.marpetgeo.2020.104422 (2020).
17. Eaton, B. A. The effect of overburden stress on geopressure prediction from well logs. *J. Pet. Technol.* **24**, 929–934. https://doi.org/10.2118/3719-PA (1972).
18. Eaton, B. A. The equation for geopressure prediction from well logs. *J. OnePetro.* **1**, 5544. https://doi.org/10.2118/5544-MS (1975).
19. Kablan, O. & Chen, T. Shale gas reservoir pore pressure prediction: A case study of the Wufeng-Longmaxi formations in Sichuan Basin, Southwest China. *J. Energies* **16**, 7280. https://doi.org/10.3390/en16217280 (2023).
20. Allawi, R. H. & Al-Jawad, M. S. Prediction of pore and fracture pressure using well logs in Mishrif reservoir in an Iraqi oilfield. *J. AIP Conf. Proc.* **2651**, 70005. https://doi.org/10.1063/5.0129689 (2023).
21. Bowers, G. L. Pore pressure estimation from velocity data: Accounting for overpressure mechanisms besides undercompaction. *J. SPE Drill. Complet.* **10**, 89–95. https://doi.org/10.2118/27488-PA (1995).
22. Bowers, G. L. & Katsube, T. The role of shale pore structure on the sensitivity of wire-line logs to overpressure. *J. AAPG Mem.* **2001**, 43–60. https://doi.org/10.1306/M76870C5 (2001).
23. Bowers, G. L. Determining an appropriate pore-pressure estimation strategy. *J. OnePetro.* **2001**, 130421. https://doi.org/10.4043/13042-MS (2001).
24. Li, Z. et al. Prediction of abnormal pressure in Lingshui deep-water high temperature formation based on improved Bowers method. *J. Xi'an Shiyou Univ. (Nat. Sci. Ed.)* **34**, 60–66. https://doi.org/10.3969/j.issn.1673-064X.2019.06.011 (2019).
25. Li, H. et al. A comprehensive prediction method for pore pressure in abnormally high-pressure blocks based on machine learning. *J. Process.* **11**, 2603. https://doi.org/10.3390/pr11092603 (2023).
26. Fan, H. A simple pore pressure estimation method for a disequilibrium compaction shale using sonic velocity. *J. Pet. Drill. Tech.* **29**, 9–11. https://doi.org/10.3969/j.issn.1001-0890.2001.05.002 (2001).
27. Ye, Z. et al. Investigation and application of a discrimination method for abnormal high formation pressure forming mechanism. *J. China Univ. Pet. (Ed. Nat. Sci.)* **36**, 102–107. https://doi.org/10.3969/j.issn.1673-5005.2012.03.017 (2012).
28. Wu, S. & Zhao, L. Discussion on the influencing factors of predicting formation pressure using Dc index method. *J. West China Explor. Eng.* **25**, 46–48. https://doi.org/10.3969/j.issn.1004-5716.2013.12.015 (2013).
29. Me, C. et al. Improvement and application of dc index method for formation pressure monitoring while drilling in Bohai Oilfield. *J. Mud Logging Eng.* **29**, 29–31. https://doi.org/10.3969/j.issn.1672-9803.2018.04.006 (2019).
30. Azadpour, M., Manaman, N. S., Kadkhodaie-Ilkhchi, A. & Sedghipour, M. Pore pressure prediction and modeling using well-logging data in one of the gas fields in south of Iran. *J. Pet. Sci. Eng.* **128**, 15–23. https://doi.org/10.1016/j.petrol.2015.02.022 (2015).
31. Matinkia, M. et al. A novel approach to pore pressure modeling based on conventional well logs using convolutional neural network. *J. Pet. Sci. Eng.* **211**, 110156. https://doi.org/10.1016/j.petrol.2022.110156 (2022).
32. Delavar, M. R. & Ramezanzadeh, A. Pore pressure prediction by empirical and machine learning methods using conventional and drilling logs in carbonate rocks. *J. Rock Mech. Rock Eng.* **56**, 535–564. https://doi.org/10.1007/s00603-022-03089-y (2023).
33. Farsi, M. et al. Predicting formation pore-pressure from well-log data with hybrid machine-learning optimization algorithms. *J. Nat. Resour. Res.* **30**, 3455–3481. https://doi.org/10.1007/s11053-021-09852-2 (2021).
34. Ahmed, A. et al. New model for pore pressure prediction while drilling using artificial neural networks. *J. Arab. J. Sci. Eng.* **44**, 6079–6088. https://doi.org/10.1007/s13369-018-3574-7 (2019).
35. Huang, H. et al. Research on prediction methods of formation pore pressure based on machine learning. *J. Energy Sci. Eng.* **10**, 1886–1901. https://doi.org/10.1002/ese3.1112 (2022).
36. Ahmed, A., Elkatatny, S., Ali, A. & Abdulraheem, E. Comparative analysis of artificial intelligence techniques for formation pressure prediction while drilling. *J. Arab. J. Geosci.* **12**, 1–13. https://doi.org/10.1007/s12517-019-4800-7 (2019).
37. Zhan, W. & Ping, C. An improved data preprocessing method in dynamic measurement. *J. Chin. J. Sens. Actuators* **23**, 558–561. https://doi.org/10.3969/j.issn.1004-1699.2010.04.022 (2010).
38. Cuptasanti, W., Torabib, F. & Saiwan, C. Modelling of crude oil bubble point pressure and bubble point oil formation volume factor using artificial neural network (ANN). *J. Chem. Eng. Trans.* **35**, 1297–1302. https://doi.org/10.3303/CET1335216 (2013).
39. Mousavirad, S. J. et al. GSK-LocS: Towards a more effective generalisation in population-based neural network training. *Alex. Eng. J.* **109**, 126–143. https://doi.org/10.1016/j.aej.2024.08.097 (2024).
40. Li, C., Hai, S. & Li, X. Prediction method of formation fracture pressure based on BP neural network optimized by genetic algorithm (GA). *J. Xi'an Shiyou Univ. (Nat. Sci. Ed.)* **30**, 75–79 (2015).
41. Deng, S. et al. A hybrid machine learning optimization algorithm for multivariable pore pressure prediction. *J. Pet. Sci.* **21**, 535–550. https://doi.org/10.1016/j.petsci.2023.09.001 (2024).
42. Yu, H., Chen, G. & Gu, H. A machine learning methodology for multivariate pore-pressure prediction. *J. Comput. Geosci.* **143**, 104548. https://doi.org/10.1016/j.cageo.2020.104548 (2020).
43. Mustafa, M. R., Rezaur, R. B., Rahardjo, H. & Isa, M. H. Prediction of pore-water pressure using radial basis function neural network. *J. Eng. Geol.* **135**, 40–47. https://doi.org/10.1016/j.enggeo.2012.02.008 (2012).
44. Liao, G. et al. Prediction of microscopic pore structure of tight reservoirs using convolutional neural network model. *J. Pet. Sci. Bull.* **5**, 26–38 (2020).

Author contributions

L.J., L.M. and L.W. proposed the innovative ideas and methods; S.B. performed the analysis, analyzed the data and wrote the paper; Y.C., H.Z. and Y.C. provided analysis data sources and made modifications to the overall ideas and quality of the paper. The authors have read and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-07048-9.

Correspondence and requests for materials should be addressed to B.S.

Reprints and permissions information is available at www.nature.com/reprints.

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