

## Original Paper

# A novel method for predicting formation pore pressure ahead of the drill bit by embedding petrophysical theory into machine learning based on seismic and logging-while-drilling data

Xu-Yue Chen<sup>a, b, \*</sup>, Cheng-Kai Weng<sup>a, \*\*,</sup>, Lin Tao<sup>c</sup>, Jin Yang<sup>a, b</sup>, De-Li Gao<sup>a, b</sup>, Jun Li<sup>a, d</sup>

<sup>a</sup> MOE Key Laboratory of Petroleum Engineering, China University of Petroleum, Beijing, 102249, China

<sup>b</sup> Hainan Institute of China University of Petroleum-Beijing, Sanya, 572000, Hainan, China

<sup>c</sup> Tianjin Branch of CNOOC Ltd., Tianjin, 300459, China

<sup>d</sup> China University of Petroleum-Beijing at Karamay, Karamay, 834000, Xinjiang, China

## ARTICLE INFO

## Article history:

Received 18 September 2024

Received in revised form

15 April 2025

Accepted 15 April 2025

Available online 17 April 2025

Edited by Jia-Jia Fei

## Keywords:

Formation pore pressure

Prediction ahead of the drill bit

Seismic and logging-while-drilling data

Machine learning

Model interpretation

## ABSTRACT

Formation pore pressure is the foundation of well plan, and it is related to the safety and efficiency of drilling operations in oil and gas development. However, the traditional method for predicting formation pore pressure involves applying post-drilling measurement data from nearby wells to the target well, which may not accurately reflect the formation pore pressure of the target well. In this paper, a novel method for predicting formation pore pressure ahead of the drill bit by embedding petrophysical theory into machine learning based on seismic and logging-while-drilling (LWD) data was proposed. Gated recurrent unit (GRU) and long short-term memory (LSTM) models were developed and validated using data from three wells in the Bohai Oilfield, and the Shapley additive explanations (SHAP) were utilized to visualize and interpret the models proposed in this study, thereby providing valuable insights into the relative importance and impact of input features. The results show that among the eight models trained in this study, almost all model prediction errors converge to 0.05 g/cm<sup>3</sup>, with the largest root mean square error (RMSE) being 0.03072 and the smallest RMSE being 0.008964. Moreover, continuously updating the model with the increasing training data during drilling operations can further improve accuracy. Compared to other approaches, this study accurately and precisely depicts formation pore pressure, while SHAP analysis guides effective model refinement and feature engineering strategies. This work underscores the potential of integrating advanced machine learning techniques with domain-specific knowledge to enhance predictive accuracy for petroleum engineering applications.

© 2025 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

Formation pore pressure represents the pressure exerted by fluids within the pores and fractures of geological formations. It serves as a crucial parameter for wellbore structural design and drilling fluid density planning, thus forming a core component of drilling operations (Deng et al., 2024; Zhang, 2011). However, the prevalent approach to predicting formation pore pressure involves applying formation pore pressure calculations derived from logging data of nearby wells directly to the design and drilling of the target

well. These methods often fail to accurately reflect the true formation pore pressure conditions of the target well (Das and Chatterjee, 2018). Therefore, achieving a precise and accurate description of formation pore pressure in the target well is of paramount importance.

Logging data provides abundant information about formation and rock properties, including formation rock density, acoustic velocity, and porosity. Theoretical methods derived from effective stress laws by Biot (1941) and Terzaghi (1943) enable the calculation of formation pore pressure using logging data. Hottmann and Johnson (1965) proposed classic theoretical methods for formation pore pressure prediction based on sonic travel time and resistivity data to characterize shale properties. Deviations from expected trends in calculated results can indicate anomalous formation pore pressure conditions. Eaton (1972, 1975) introduced significant theoretical equations utilizing resistivity data and a new

\* Corresponding author.

\*\* Corresponding author.

E-mail addresses: [chenxuyue@cup.edu.cn](mailto:chenxuyue@cup.edu.cn) (X.-Y. Chen), [weng\\_chengkai@163.com](mailto:weng_chengkai@163.com) (C.-K. Weng).

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

equation using sonic differential time data to predict formation pore pressure (Eaton method). The empirical constant in the Eaton method needs to be determined within the range of 0.6–3 and varies in different regions. Bowers (1995) introduced a new formation pore pressure prediction method in 1995 (Bowers method) considering rheological differences through loading and unloading divisions, although it may overestimate formation pore pressure in formations with slow acoustic velocities (Zhou et al., 2020). Subsequently, various scholars have refined the Eaton and Bowers methods from multiple perspectives (Gholilou et al., 2017; Guo et al., 2023; Sun et al., 2024; Zhang, 2011). Over the past few decades, these theoretical models have provided accurate predictions and inversion of formation pore pressure in petroleum drilling. However, most of these models are empirical and single-variable, thereby exhibiting lower stability (Wang and Wang, 2015).

In recent years, machine learning has been applied to various aspects of petroleum drilling, including rate of penetration (ROP) prediction, productivity evaluation, formation pore pressure prediction, kick detection, and downhole complex condition recognition (Chen et al., 2023, 2024; Li et al., 2022a; Mahmoud et al., 2024; Wang et al., 2023). Machine learning models offer solutions to the drawbacks of single-variable and empirical parameters in traditional theoretical models. Previous studies employing machine learning for formation pore pressure prediction have predominantly focused on data-driven approaches (Li et al., 2022a). For example, researchers have utilized logging data for formation pore pressure prediction (Lockhart et al., 2023). Ahmed et al. (2019) creatively applied support vector machine (SVM) to predict pore and fracture pressures with high accuracy based on real field data, outperforming traditional models in simplicity and prediction ability. Xu et al. (2024) utilized parameters such as depth, spontaneous potential, natural gamma ray, and sonic differential time from logging data as inputs to build and compare back propagation neural network (BPNN), recurrent neural network (RNN), and LSTM neural networks. They enhanced model applicability using LSTM combined with transfer learning algorithms. The experimental results showed that compared with the LSTM model, the TCA-LSTM model using transfer component analysis had a better prediction effect, with an average absolute error of only 0.50%. Some researchers have predicted and monitored formation pore pressure using surface logging data (Li et al., 2023). Ahmed et al. (2021) developed an artificial neural network using parameters recorded at the wellhead, including rate of penetration, mud flow rate, standpipe pressure, and rotary speed, for real-time monitoring of formation pore pressure changes. The model achieved an  $R^2$  of 0.98, indicating its effectiveness in accurately inverting formation pore pressure at the wellhead. Others have employed seismic data for formation pore pressure prediction (Lockhart et al., 2023). Li et al. (2023) introduced a unique deep learning model, TGG, integrating temporal convolution, graph adaptive learning, and graph convolution, which demonstrated remarkable performance in predicting abnormal pore pressure with high accuracy and robustness using field kick data. Zhang et al. (2024) introduced a highly generalized CGP-NN model that utilized seismic attributes such as instantaneous frequency, instantaneous  $Q$  related to energy attenuation, and maximum curvature to train a multi-layer GRU model. The model achieved an  $R^2$  of 0.94 on the validation set, surpassing the performance of traditional methods like Eaton and Bowers. Furthermore, some researchers have integrated domain knowledge into machine learning models for formation pore pressure prediction. Cao et al. (2024) proposed the knowledge-aware TFT Model, a new deep learning framework that combines geological knowledge with formation pore pressure evolution theory to achieve high-precision formation pore pressure prediction.

The aforementioned scholars have expanded the range of parameters used for predicting formation pore pressure and have trained models that achieve relatively high accuracy. However, many models developed in these studies using logging data appear more like inversion models for formation pore pressure after drilling rather than predictive models aimed at accurately characterizing formation pore pressure in target wells (Wang and Wang, 2015). Data collected at the wellhead often suffers from latency issues and cannot provide real-time reflections of conditions at the bottom of the well (Jorden and Shirley, 1966). While using seismic attribute parameters can address the challenge of predicting formation pore pressure before drilling into target wells, acquiring multiple seismic attribute parameters is difficult, and there currently lacks mature theory and practical guidelines demonstrating that using seismic attribute parameters alone can finely describe formation pore pressure. Moreover, previous machine learning models established by researchers are often black-box models that do not provide insights into the contributions of each parameter in predicting formation pore pressure.

Currently, with the advancement of LWD technology, some logging tools can achieve ahead-of-the-bit detection, obtaining formation property parameters from several tens of meters deep beneath the drill bit (Guo et al., 2020; Hagiwara, 2018; Khalil et al., 2018; Wang et al., 2022). These tools enable high-precision imaging, long-range detection, proactive sensing, passive integrated multi-parameter capabilities, and real-time formation testing during drilling (Wang and Ye, 2024). Building upon the shortcomings identified by previous scholars and the development of logging tools, this study proposes a novel approach to accurately predict formation pore pressure by integrating petrophysical theory into a machine learning model using fused seismic and LWD data from three deep-sea wells in the Bohai Oilfield. This method achieves precise prediction of formation pore pressure beneath the drill bit, addressing challenges in traditional methods and previous research by accurately characterizing formation pore pressure in target wells and providing interpretability of the machine learning models.

## 2. Geological background and data processing

All data in this study were sourced from the Bohai Oilfield, China. The complex geological structures and sedimentary backgrounds pose challenges for accurate prediction of formation pore pressure, and insights gained from drilling complexities in historical wells are beneficial for assessing the distribution of formation pore pressure in the study area. Therefore, understanding the relevant geological background and certain historical drilling incidents is crucial for data selection and model training. Wells A, B, and C are all vertical wells drilled within the past five years.

### 2.1. Historical drilling complications

The study area features an overall buried-hill structural background, with the basement developed in the Archean, Paleozoic, and Mesozoic eras (Xue and Wang, 2020). The strata are divided into the Pingyuan Formation (Qp), Minghuazhen Formation ( $N_2m$ ), Guantao Formation ( $N_1g$ ), Dongying Formation ( $E_3d$ ), Shahejie Formation ( $E_{2-3s}$ ), Mesozoic (Mz), Paleozoic (Pz), and Archean (Ar) from top to bottom. The Dongying Formation is further subdivided into three members, Dong 1 Member, Dong 2 Member (including three submembers of  $E_{3d2 I}$  and  $E_{3d2 II}$ ) and Dong 3 Member. Historical drilling instances demonstrate abnormal overpressure in the Dong 3 Member to the Shahejie Formation (Liu et al., 2019; Pu et al., 2020). The statistics of the complex situation of well kick (Kick) and drilling fluid loss (Loss) in some adjacent wells are shown in Table 1.

**Table 1**  
Statistics of complex accidents in adjacent wells.

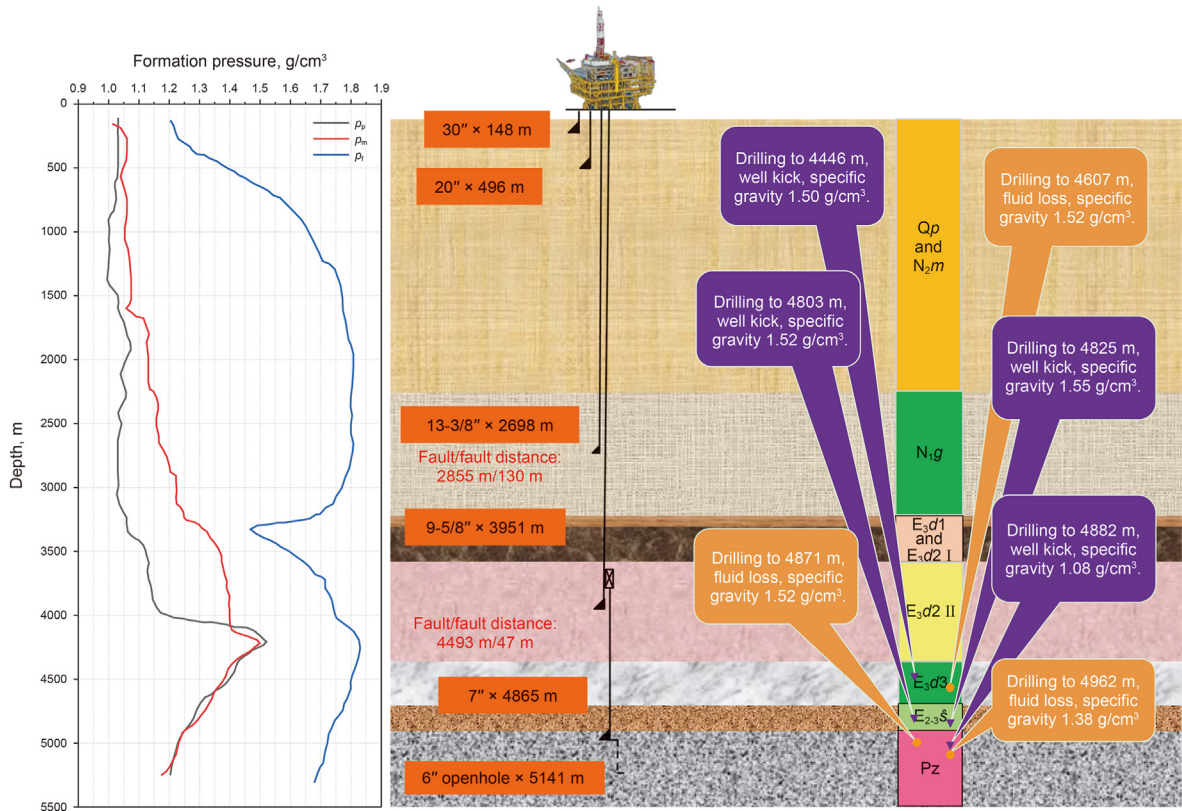
| Well | Wellbore size, in | Depth, m | Complex situations | Measure                                                         | Stratigraphic level |
|------|-------------------|----------|--------------------|-----------------------------------------------------------------|---------------------|
| Z-1  | 8-1/2             | 4446     | Kick               | Cycle to increase specific gravity to 1.52 g/cm <sup>3</sup> .  | Dong 3 Member       |
|      | 8-1/2             | 4607     | Loss               | Reduce displacement to 1600 L/min.                              | Shahejie Formation  |
|      | 8-1/2             | 4803     | Kick               | Normal after cyclic exhaust.                                    |                     |
|      | 8-1/2             | 4825     | Kick               | Increase the drilling fluid density to 1.55 g/cm <sup>3</sup> . | Paleozoic           |
|      | 8-1/2             | 4871     | Loss               | Reduce the displacement to 400 L/min.                           |                     |
|      | 6                 | 4882     | Kick               | Increase the drilling fluid density to 1.33 g/cm <sup>3</sup> . |                     |
|      | 6                 | 4962     | Loss               | Add plugging material.                                          |                     |
| Z-2  | 12-1/4            | 3963     | Loss               | Reduce the displacement to 3000 L/min.                          | Dong 2 Member       |
|      | 6                 | 5224     | Loss               | —                                                               | Archean             |
| Z-3  | 6                 | 4378     | Kick               | Increase the drilling fluid density to 1.20 g/cm <sup>3</sup> . | Paleozoic           |
| Z-4  | 12-1/4            | 4223     | Loss               | —                                                               | Dong 3 Member       |
|      | 12-1/4            | 4212     | Kick               | —                                                               |                     |
|      | 12-1/4            | 4328     | Loss               | —                                                               |                     |
|      | 12-1/4            | 4438     | Loss               | Add plugging material with drilling.                            |                     |
|      | 8-1/2             | 4645     | Kick               | —                                                               | Mesozoic            |

During drilling of Well Z-1, there were recorded 7 instances of complex drilling fluid losses and well kicks. Fig. 1 illustrates the wellbore design guided by pre-drill predictions of the three-pressure profiles for Z-1, juxtaposed with the complexities encountered during actual drilling. Specific analysis was conducted on the application of pre-drill predictions of formation pore pressure in the Z-1 well during its drilling in the Bohai Oilfield, highlighting shortcomings of the traditional Eaton method. Both the predicted three-pressure profiles and the drilling complexities indicated an initial rise in formation pore pressure in the lower section of Dong 2 Member. The predicted profiles suggested a subsequent decline in formation pore pressure towards the base of Dong 2 Member, yet in actual drilling, elevated pressures persisted within the Shahejie Formation, reaching densities of 1.52–1.55 g/cm<sup>3</sup>.

This indicates that there is a continuous high-pressure interval encountered in the formations drilled by Well Z-1 from the lower section of Dong 2 Member down to the base of Shahejie Formation, rather than the predicted scenario where pressure was expected to peak at the base of Dong 2 Member and then decline. It also suggests that regressing well logging data obtained from neighboring wells post-drilling cannot directly refine the description of formation pore pressure in the target well.

2.2. Raw data analysis

Fig. 2 illustrates the relative positions of wells A, B, and C. Wells B and C are situated within the same block, while Well A is located in a different block. Fig. 3 depicts the variation of seismic velocity-



**Fig. 1.** Design of well structure guided by pre-drilling prediction of three pressure profiles in Well Z-1 and complications arising from the actual drilling.



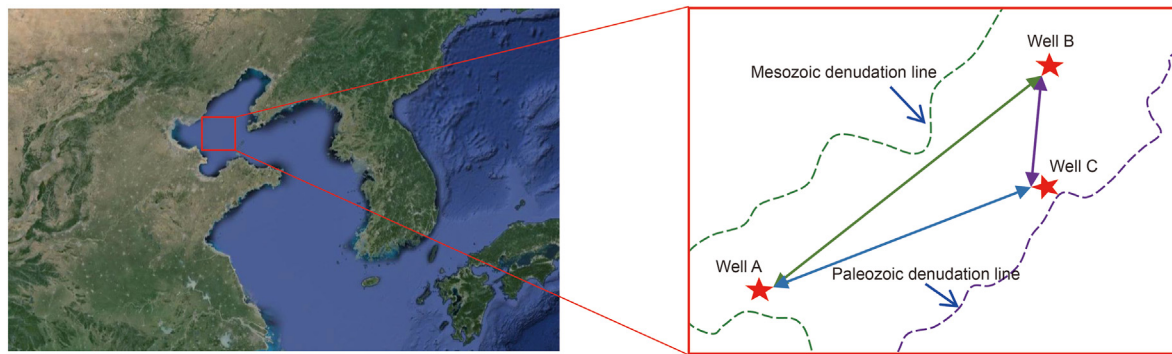


Fig. 2. Position of wells A, B, and C.

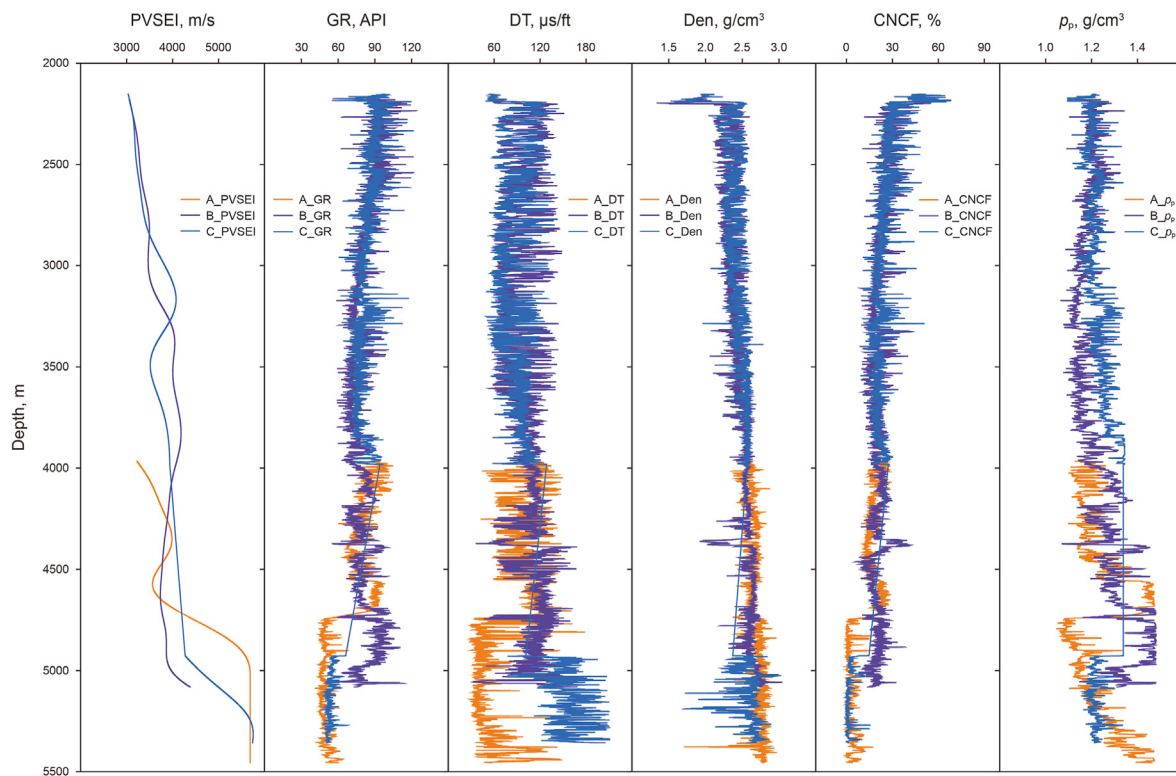


Fig. 3. Original data and corrected formation pore pressure.

related parameter (seismic layer velocities (PVSEI)) and four logging-related parameters (natural gamma ray (GR); sonic differential time (DT); rock density (Den); porosity (CNCF)) with depth for these three wells. From the figure, it is evident that wells B and C, which are in the same block, exhibit similar trends in formation pore pressure, whereas Well A from the different block shows dissimilar trends compared to wells B and C.

From Fig. 3, significant variations in PVSEI among the three wells are observed. Wells B and C exhibit similar changes in DT, Den, and CNCF in the shallow formations. In terms of GR, notable differences between these two wells begin around 3068 m, aligning with their PVSEI variations. It is evident that before 2800 m, the trends in formation pore pressure for both wells are nearly identical, mirroring the changes in PVSEI. However, at 3150 m, a slight increase in formation pore pressure is observed in Well C, while a slight decrease is seen in Well B. This significant trend change is minimal reflected in GR, DT, Den, and CNCF, though GR and CNCF show numerical changes, these are marginal and could easily be dismissed as data recording errors. In contrast, PVSEI clearly indicates a minor decline

in Well B and a slight increase in Well C. Furthermore, below 4500 m, significant fluctuations in GR, DT, CNCF, and PVSEI values distinctly capture the sharp decrease followed by a rapid increase in formation pore pressure. This highlights the seismic parameters' ability to effectively and intuitively characterize formation pore pressure variations with depth. This reaffirms the effectiveness of integrating multiple sources, including seismic and LWD parameters, for formation pore pressure prediction in this study.

### 2.3. Data processing

To effectively demonstrate the effectiveness of this method, this study used data from the upper well sections of wells A, B, and C as both training and testing sets to establish models. These models were then used to predict formation pore pressure in the lower well sections of each respective well. Directly inputting data into the model could lead to issues such as large numbers dominating small ones, potentially causing the model to overlook certain parameters (Chen et al., 2023). Therefore, Min-Max normalization was employed to

scale all data into the  $[-1, 1]$  range. This normalization aids in faster convergence of optimization algorithms and enhances model accuracy. The formula for Min-Max normalization is:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

In the formula,  $x$  is a certain set of feature data, and  $x'$  is the data after the change,  $\max$  and  $\min$  indicate the maximum and minimum values of the data, respectively.

The interpolation and supplementation of data aim to retain as many data points as possible while introducing a limited and acceptable level of error. The purpose of this operation is to reduce the impact of anomalous data on model convergence and to prevent errors during model training. When handling anomalies in the data, Kriging interpolation is used to fill in the missing values. Kriging interpolation combines the data from known locations using weights to estimate the values at missing points. For the missing value at  $x_i$ , it can be expressed as:

$$Y(x_i) = \sum_{j=1}^n [\lambda_j * Y(x_j)] \quad (2)$$

where,  $Y(x_i)$  is the value at the missing point;  $Y(x_j)$  is the value at the known point;  $n$  is the number of known points involved in the interpolation;  $\lambda_j$  is the weight for each known point. The core of Kriging interpolation is to determine the weights that minimize bias and ensure the smallest variance in the interpolation results. The calculation of the weights depends on the spatial correlation between the points, which can be measured using a semivariogram. The definition of the semivariogram is as follows:

$$\gamma(h) = \frac{1}{2} E[(Y(x) - Y(x+h))^2] \quad (3)$$

where,  $h$  is the distance between two points;  $Y(x)$  and  $Y(x+h)$  are the measurements at locations  $x$  and  $x+h$ , respectively.  $E$  is the mathematical expectation measuring the average of  $[(Y(x) - Y(x+h))^2]$ . The weights  $\lambda_j$  can be determined by solving the following system of linear equations:

$$\sum_{j=1}^n \lambda_j \gamma(x_i - x_j) + \mu = \gamma(x_i - x_j), \quad k = 1, 2, \dots, n \quad (4)$$

$$\sum_{j=1}^n \lambda_j = 1 \quad (5)$$

where  $\mu$  is the Lagrangian residual constrained to ensure the unbiasedness of the weight. After calculating the semivariogram  $\gamma(x_i - x_j)$  between each point using the above Eq. (10), each  $\lambda_j$  is determined by solving the linear equation system determined by Eqs. (11) and (12). Finally,  $\lambda_j$  is substituted into the Kriging interpolation formula of Eq. (9) to obtain the missing value  $Y(x_i)$  at  $x_i$ .

To further validate the relationships and dependencies among the selected variables in this study, Spearman's rank correlation coefficient was employed to analyze the correlations between various parameters. This analysis helps uncover patterns and relationships within the data. Spearman's correlation coefficient can be calculated using the following formula (He et al., 2023):

$$\rho = \frac{\frac{1}{n} \sum_{i=1}^n (R(x_i) - \bar{R}(x))(R(y_i) - \bar{R}(y))}{\sqrt{\left(\frac{1}{n} \sum_{i=1}^n (R(x_i) - \bar{R}(x))^2\right) \cdot \left(\frac{1}{n} \sum_{i=1}^n (R(y_i) - \bar{R}(y))^2\right)}} \quad (6)$$

Here,  $R(x)$  and  $R(y)$  denote the ranks of  $x$  and  $y$ , respectively, while  $\bar{R}(x)$  and  $\bar{R}(y)$  represent the average ranks.

Fig. 4 depicts the computed Spearman correlation coefficients. From the correlation plot, Depth, GR, DT, and Den show a positive correlation with formation pore pressure, with values around 0.6, indicating a moderate correlation, which meets the requirements for machine learning input parameters. On the other hand, PVSEI exhibits a weak correlation with formation pore pressure. This is expected because the formation pore pressure data used in this study are derived from the Eaton method and adjusted during drilling operations for factors such as drilling fluid density.

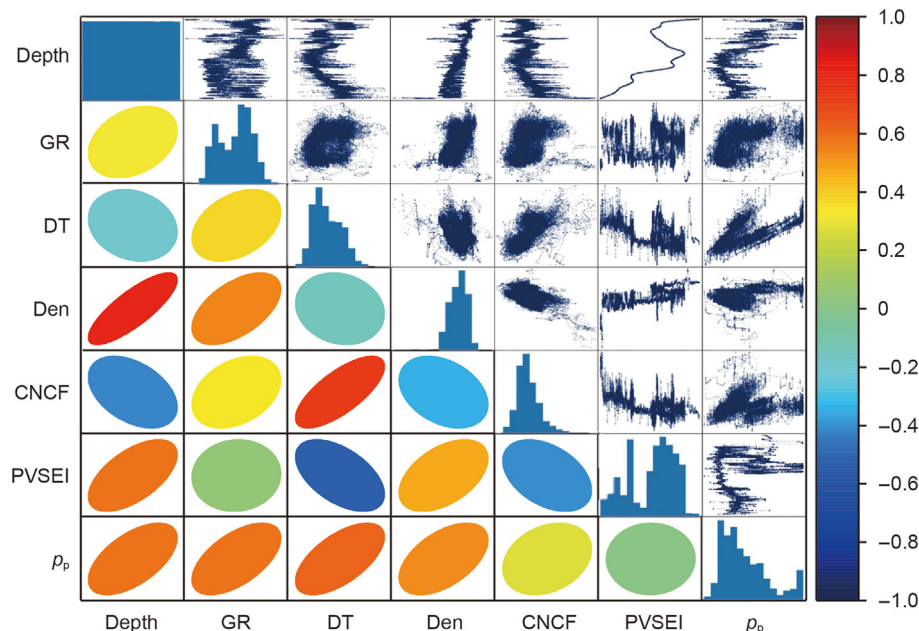


Fig. 4. Spearman's correlation coefficient.

### 3. Methodology

#### 3.1. Formation pore pressure prediction model

Hubbert and Rubey (1959) introduced the concept of effective stress into the field of geology. They established the relationship between formation pore pressure, overburden pressure, and vertical effective stress as follows:

$$p_p = p_o - \sigma \quad (7)$$

where  $p_p$  is predicted formation pore pressure,  $p_o$  is overburden pressure, and  $\sigma$  is vertical effective stress.

Based on logging data, Eaton (1972, 1975) proposed two well-known theoretical models for predicting formation pore pressure, including models based on shale resistivity and sonic differential time. According to extensive historical research, the primary mechanism for abnormal high pressure in the Bohai Oilfield, China, is attributed to under compaction of shale, consistent with the effective utilization scope of the Eaton method. Under the principle of effective stress, the Eaton method can be expressed as:

$$p_p = p_o - (p_o - p_h) \left( \frac{DT_n}{DT_o} \right)^n \quad (8)$$

where  $p_h$  is normal hydrostatic pressure,  $DT_n$  is normal sonic transit time,  $DT_o$  is logging-determined rock sonic transit time, and  $n$  is the Eaton exponent.

The Eaton method is based on the principle of effective stress and the intrinsic relationship between formation pore pressure and seismic wave propagation velocity. The mechanical and physical properties of rocks are significantly influenced by effective stress. Effective stress refers to the stress acting on the rock framework, and it has a specific relationship with the total stress and pore pressure. In the subsurface geological environment, total stress can be considered as the overburden pressure, which is borne jointly by the rock framework and pore fluids. When pore pressure changes, effective stress also undergoes modifications, thereby affecting the rock's porosity structure and other physical properties. When pore pressure causes structural changes in the rock, the contact relationships between the rock particles also change, leading to a reduction in the rock's stiffness. According to the theory of elastic wave propagation, the speed of seismic waves in rock is closely related to the rock's stiffness, and a decrease in stiffness results in a reduction in wave propagation speed. The Eaton method effectively applies this physical principle to correlate seismic wave velocity with pore pressure.

#### 3.2. Physical mechanisms embedded in machine learning

To enable machine learning models to better capture natural laws, some scholars have integrated physical theories into machine learning models by modifying the loss function (Bommidi et al., 2023; Ribeiro et al., 2024; Loutfi et al., 2022; Xiong et al., 2024). Studies have shown that this approach effectively enhances the predictive efficiency and stability of models. The loss function measures the difference between model predictions and actual values, serving as the core of optimization algorithms. Different loss functions are suitable for different problems and data types, leading to different models being trained. In regression problems, Mean

Squared Error (MSE) is commonly used as a loss function, defined as:

$$MSE = \frac{1}{n} \sum_{k=1}^n (y_{tk} - y_{pk})^2 \quad (9)$$

where  $y_{tk}$  is the true values on the test set and  $y_{pk}$  is the predicted values on the test set. MSE measures the squared difference between model predictions and actual values, providing a quantitative assessment of prediction accuracy.

Embedding physical theories into the loss function integrates domain knowledge into the process of training machine learning models. The primary objective is to introduce domain knowledge to enhance the applicability of the models for predicting formation pore pressure, thereby improving model performance and interpretability (Lan et al., 2023). Petrophysical theory impose additional constraints on model outputs, ensuring better alignment with physical laws governing the natural world, thereby enhancing model reliability and robustness. In traditional machine learning models, such as those mentioned above that use MSE as the loss function, the primary mechanism is to measure the deviation between the model's predicted values and the true values. The model primarily relies on the data itself for training, and in some complex problems, machine learning models may fail to capture the underlying patterns within the data. In particular, when there is significant data noise, the robustness of the model during training may not be stable. The introduction of rock physics theory can assist the model in making more reasonable predictions when faced with unseen data. Moreover, aligning the model's outcomes with established physical laws or mechanisms in the domain enhances the interpretability of integrating physical theories into machine learning models. Previous machine learning models were often considered "black-box" models, making it difficult to explain their internal working mechanisms. The introduction of rock physics theory ensures that the model's output not only aligns with the data but also adheres to physical laws.

This study implemented the calculation of formation pore pressure based on the Eaton method using Python, formulated as follows:

$$p_p = p_p\text{-Compute}(\text{Depth}, DT, \text{Den}) \quad (10)$$

where Den is rock density. Since machine learning is fundamentally data-driven, it learns relationships and patterns within data through extensive datasets. Using only physical theories as the loss function may not fully utilize the potential information within the data, thereby limiting the model's performance and generalization ability. Real-world data inherently includes noise and uncertainty, whereas theoretical mechanisms are often derived under idealized conditions. Relying solely on physical theories as the loss function could lead to overfitting. Complex deep learning models tend to perform well on the training data but poorly on the test data because the model learns noise or irrelevant features in the data. Integrating MSE with a physical theory formula as the loss function balances the influences of data and physical theories on the model. The loss function in this study is expressed as:

$$\text{Loss} = \lambda \cdot \sqrt{(p_p\text{-Compute}(\text{Depth}, DT, \text{Den}) - y_{pk})^2} + (1 - \lambda) \cdot \text{MSE} \quad (11)$$

Here,  $\lambda$  is the weight of the physical mechanism.

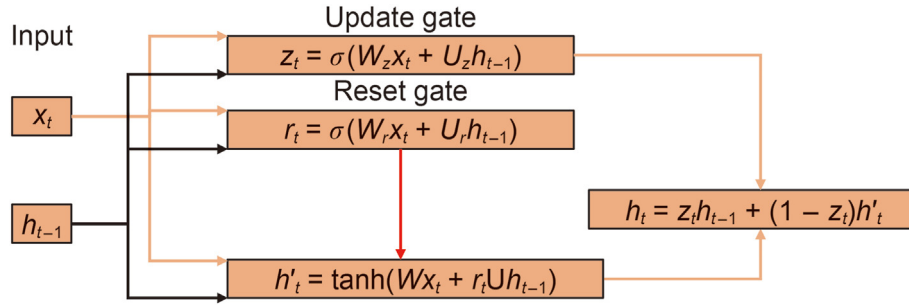


Fig. 5. GRU principle diagram.

### 3.3. Introduction to machine learning models as applied to pore pressure prediction

This study investigates two machine learning models, GRU and LSTM. GRU and LSTM are specifically designed to handle long-term dependencies in time series data. They can capture long-range temporal correlations through memory and forget mechanisms, and are suitable for complex time series tasks. During the drilling process, various parameters that vary with depth can be considered as time series data. Therefore, it is more appropriate to use GRU and LSTM. Both GRU and LSTM embed physical mechanisms into machine learning by modifying the Loss function. GRU and LSTM are common sub-models of recurrent neural networks (RNNs) used for sequential processing (Chen et al., 2023, 2024). Schematics illustrating the principles of GRU and LSTM for predicting formation pore pressure are shown in Figs. 5 and 6, respectively.

### 3.4. Shapley additive explanations of pore pressure prediction model

Understanding why machine learning models make certain predictions and analyzing the impact of parameters on these predictions is crucial. Recently, SHAP has been widely applied in various fields to interpret machine learning models (Kannangara et al., 2022). For instance, Zhao et al. (2023) used SHAP visualizations to clarify the decision-making process and interpret their tree model's predictions for formation pore pressure under high-temperature and high-pressure conditions. In their study, SHAP was employed to explain the model's behavior and assess the contribution of each parameter to the predicted pore pressure values. SHAP, grounded in game theory, is a method for explaining model outputs by assigning an importance value—known as the Shapley value—to each feature, based on its contribution to the model's predictions (Białek et al., 2022). The Shapley value, derived

from cooperative game theory, ensures a fair distribution of total rewards based on each player's contribution. SHAP calculates each feature's contribution by evaluating all possible feature combinations and their respective outputs from the machine learning model. To simplify the interpretation, SHAP maps simplified inputs back to the original input space using a function (Song et al., 2024). This explanatory model can be defined as follows:

$$e(z') = \phi_0 + \sum_{i=1}^N \phi_i z'_i \quad (12)$$

Here,  $e$  is the explanation function,  $z'$  is a subset of simplified binary features (hence  $z' \in \{0, 1\}^N$ ),  $N$  is the number of simplified features, and  $\phi_i$  is coefficients in the explanation model indicating the contribution of the  $i$ -th feature to the model prediction.  $\phi_0$  is a constant known as the expected value of the model prediction, which in SHAP equals the mean prediction of the model on the dataset. Importantly, models of this form exhibit specific properties where the sum of contributions from all important features approximately equals the output of the original model  $f$ , hence  $f(x) \approx e(x)$ . The computation for each feature's contribution is:

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(N - |z'| - 1)!}{N!} [f_x(z') - f_x(z' \setminus i)] \quad (13)$$

The contribution of feature  $i$ -th is the average difference in predictions for all simplified feature subsets that include and exclude the feature.

### 3.5. Workflow of predicting formation pore pressure ahead of the drill bit

Fig. 7 illustrates the workflow of this study. In this work, we utilized integrated seismic and LWD data and employed a machine learning model embedded with petrophysical theory as the

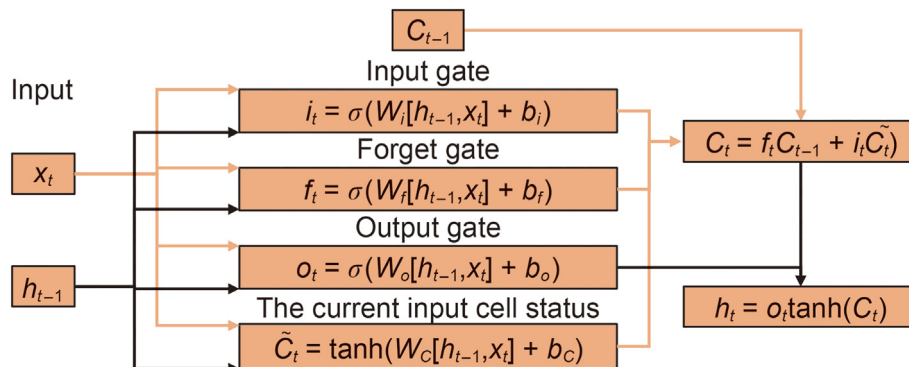


Fig. 6. LSTM principle diagram.



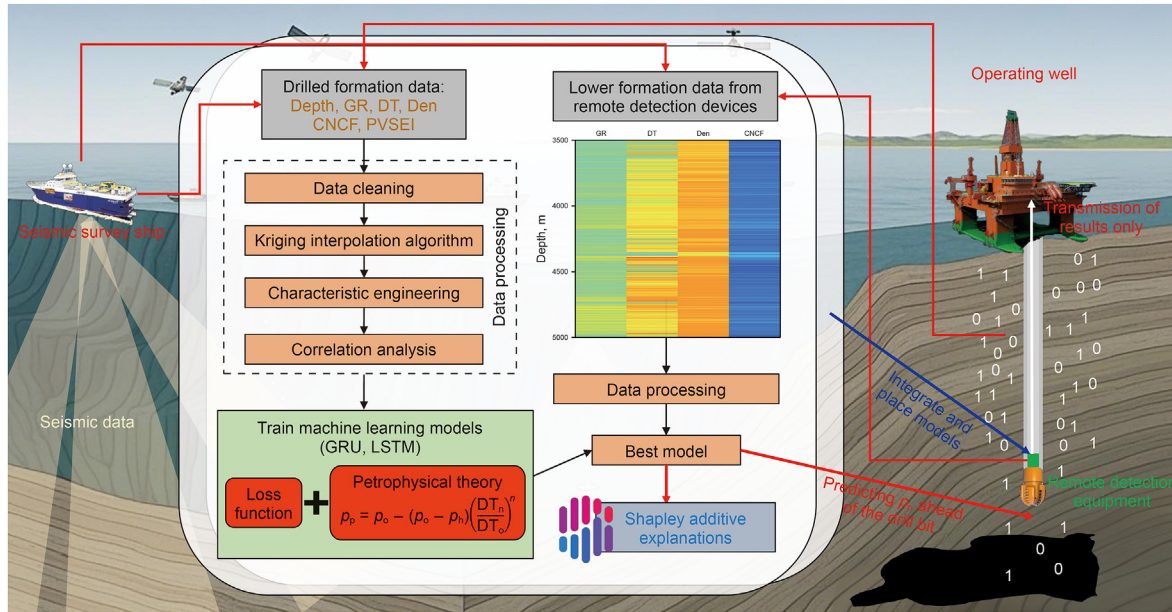


Fig. 7. Workflow of predicting formation pore pressure ahead of the drill bit.

processor. We propose a novel approach placing the integrated model downhole to predict formation pore pressure beneath the drill bit. This method involves recording seismic data of the formations, logging data from upper drilled sections, and downhole LWD data measured by the remote detection equipment (RDE). Data collection and processing are completed downhole, and results are computed in real-time, transmitted to surface equipment via mud pulse telemetry. Direct identification of formation pore pressure downhole and transmitting computed results to the surface significantly mitigates current downhole data transmission bottlenecks. To demonstrate the feasibility of this new method in model implementation, this study unfolds through model establishment and engineering application simulations.

This study unfolds through six key steps of the proposed method:

**Step 1: Data collection.** It is necessary to extract pre-drilling seismic exploration data from the block and logging data from the drilled sections of the operational wells, along with formation physical property data, for the inversion of formation pore pressure. Formation pore pressure of the drilled sections is computed using physical method such as Eaton method, and adjusted for encountered conditions during drilling, such as drilling fluid density, formation physical properties, and complexities conditions during drilling, serving as the result labels.

**Step 2: Data processing.** The step intervals recorded in seismic and logging data are inconsistent, necessitating data normalization. Occasional data gaps due to downhole complexities, instrument issues, or data recording require interpolation using Kriging. Proper partitioning of training data into testing and validation sets enhances the predictive model's generalization capabilities effectively (Chen et al., 2023; Ma et al., 2024).

**Step 3: Embedding petrophysical mechanisms.** Embedding physical theories into machine learning models incorporates natural world operating principles beyond mere data-driven operation, detailed in Section 1.3.

**Step 4: Training formation pore pressure prediction models.** Models are trained using data from drilled well sections with

two machine learning approaches. Different machine learning mechanisms, such as GRU and LSTM models selected in this study, may perform differently in various applications.

**Step 5: Engineering application.** During the drilling process of operational wells, seismic data for the operational well is known, while downhole formation data is sourced from RDE. Cleaning and passing seismic data and ahead-of-the-bit LWD data to the selected formation pore pressure prediction model enable precise prediction of formation pore pressure. Continuous updating of the model with data from the target well during drilling improves prediction accuracy continually.

**Step 6: Post-drilling model interpretation.** Utilizing SHAP, the predictions of the applied machine learning models across all depths of the operational well are interpreted. This analysis assesses the relative importance and impact of each feature in the model, guiding model improvements, feature selection, or engineering.

### 3.6. Model evaluation criteria

In order to assess the accuracy of the model of this study in many aspects, this paper uses six evaluation models to evaluate the effectiveness of the model trained in this study in each well. Each evaluation method is calculated as follows (Zhang et al., 2023):

(1) Root mean square error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{k=1}^n (y_{tk} - y_{pk})^2} \quad (14)$$

(2) Mean absolute error (MAE):

$$\text{MAE} = \frac{1}{n} \sum_{k=1}^n |y_{tk} - y_{pk}| \quad (15)$$



**Table 2**  
Hyperparameter usage of GRU and LSTM.

| Hyperparameters      | GRU  | LSTM | Hyperparameters   | GRU                          | LSTM |
|----------------------|------|------|-------------------|------------------------------|------|
| Test set ratio       | 0.3  |      | Loss function     | Weight of MSE and Eaton loss |      |
| Random seed          | 42   |      | MSE weight        | 0.5                          |      |
| First layer neurons  | 64   |      | Eaton loss weight | 0.5                          |      |
| Second layer neurons | 32   |      | Number of epochs  | 100                          |      |
| Output layer neurons | 1    |      | Batch size        | 32                           |      |
| Optimizer            | Adam |      | L1 regularization | 0.01                         |      |

(3) Mean absolute percentage error (MAPE):

$$MAPE = \frac{1}{n} \sum_{k=1}^n \left| \frac{y_{tk} - y_{pk}}{y_{tk}} \right| \quad (16)$$

(4) Coefficient of determination ( $R^2$ ):

$$R^2 = 1 - \frac{\sum_{k=1}^n (y_{tk} - y_{pk})^2}{\sum_{k=1}^n (y_{tk} - \bar{y})^2} \quad (17)$$

(5) Percent deviation between measured and predicted values for the  $k$ -th dataset record ( $PD_k$ ):

$$PD_k = \frac{y_{pk} - y_{tk}}{y_{pk}} \times 100\% \quad (18)$$

(6) Absolute average percent deviation (AAPD):

$$AAPD = \frac{\sum_{k=1}^n |PD_k|}{n} \quad (19)$$

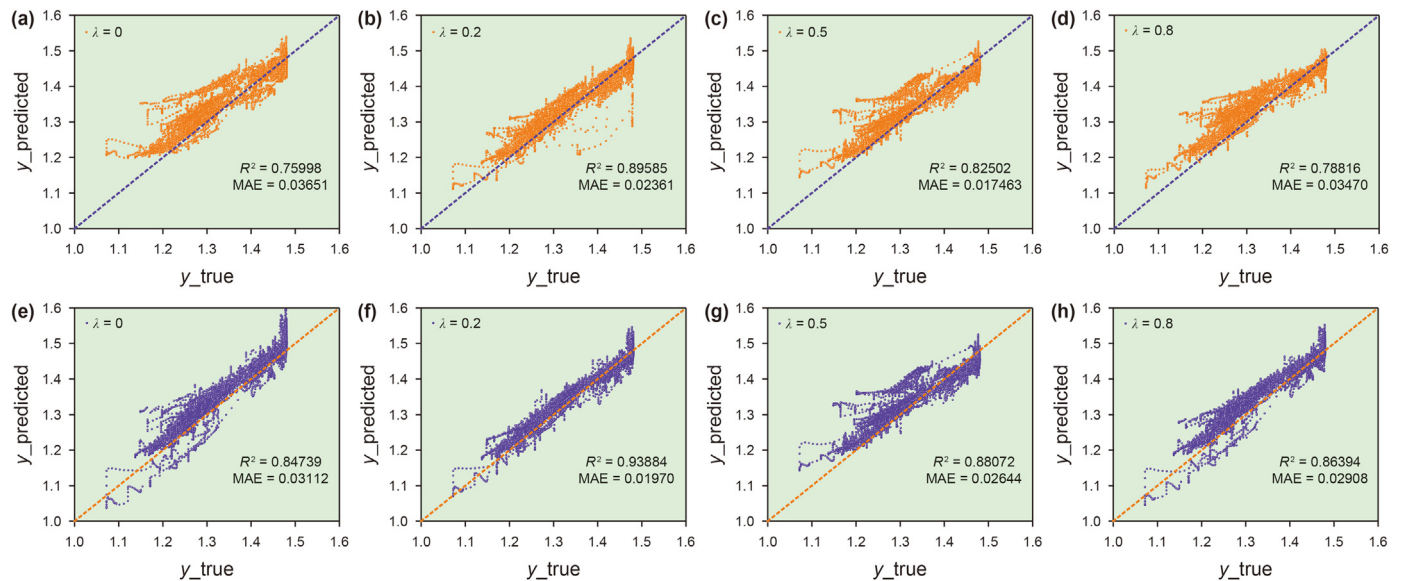
## 4. Results analysis and discussion

### 4.1. Choice of model hyperparameters and physical weights

In order to reflect the effectiveness of the case study, all GRU and LSTM models in this study use the same and simple hyperparameters, but the loss function is not the traditional MSE evaluation method. Table 2 lists the hyperparameter usage of GRU and LSTM.

For a specific block and method, determining an appropriate physical weight, is a crucial foundation for building a pore pressure prediction model. In this study, we set physical weights to (0, 0.2, 0.5, 0.8) for iterative training. When  $\lambda = 0$ , it indicates the absence of physical constraints, making the model a purely data-driven pore pressure prediction model. In this study, we used the first 20000 records of Well B as training sets and validation sets to train the model, and the lower 9032 records as the test set for engineering application simulation. The model performance under different physical weights is analyzed to determine the optimal  $\lambda$  for establishing the model in this article.

Fig. 8 presents a comparative analysis of true and predicted values using GRU and LSTM models under varying physical weights. The results indicate that incorporating physical constraints significantly impacts the model's predictive ability. When  $\lambda = 0$ , the pure data-driven approach shows relatively poor performance in both GRU and LSTM models (GRU  $R^2 = 0.75998$ , LSTM  $R^2 = 0.84719$ ). This suggests that data-driven learning alone may not fully capture the complex physical underpinnings of pore pressure. Notably, both models achieve optimal performance when  $\lambda = 0.2$ , with the LSTM model ( $R^2 = 0.93884$ ) outperforming the



**Fig. 8.** Performance comparison of models built with different physical weights in GRU and LSTM frameworks.

GRU model ( $R^2 = 0.89585$ ). This optimal weight represents a well-balanced integration of physical constraints and data-driven learning, effectively preventing underfitting and overfitting in the model. The enhanced performance of the LSTM model can be attributed to its superior ability to capture complex sequential patterns in data with long-term dependencies. However, increasing the physical weight may lead to a decline in model performance. In this study, the performance of models with  $\lambda = 0.5$  and  $\lambda = 0.8$  weights was lower than that of the model with  $\lambda = 0.2$ . This suggests that an overemphasis on physical constraints may weaken the model's ability to learn from actual data patterns, particularly in regions where physical theories cannot fully capture or describe the complexities of local geology.

#### 4.2. Results and evaluation

In this study, Well A had a dataset of 14453 records, with the first 10000 records (up to a depth of 4996.1 m) used for training and

testing. Well B had a dataset of 29032 records, with the first 10000 (up to a depth of 3176.6 m) and the first 20000 records (up to a depth of 4176.6 m) used separately for training and testing. Well C had a dataset of 22352 records, with the first 20000 records (up to a depth of 5097.3 m) used for training and testing. The proposed method trained GRU and LSTM models for each well, and the prediction results are shown in Fig. 9. The yellow boxes in the figure represent the training and validation sets, while the red boxes indicate the application of models trained using upper formation data to predict lower formation pressures. Figs. 10 and 11 depict the errors and distributions of the GRU and LSTM model predictions compared with post-drilling corrections across the three wells.

The trends in results obtained from applying both machine learning frameworks in wells A, B, and C generally align with post-drilling corrections. For instance, at 5316 m in Well A, GR, DT, DEN, and CNCF all exhibit significant trends, with both models identifying a minor sharp rise in formation pore pressure. This consistency indicates that both machine learning models incorporating

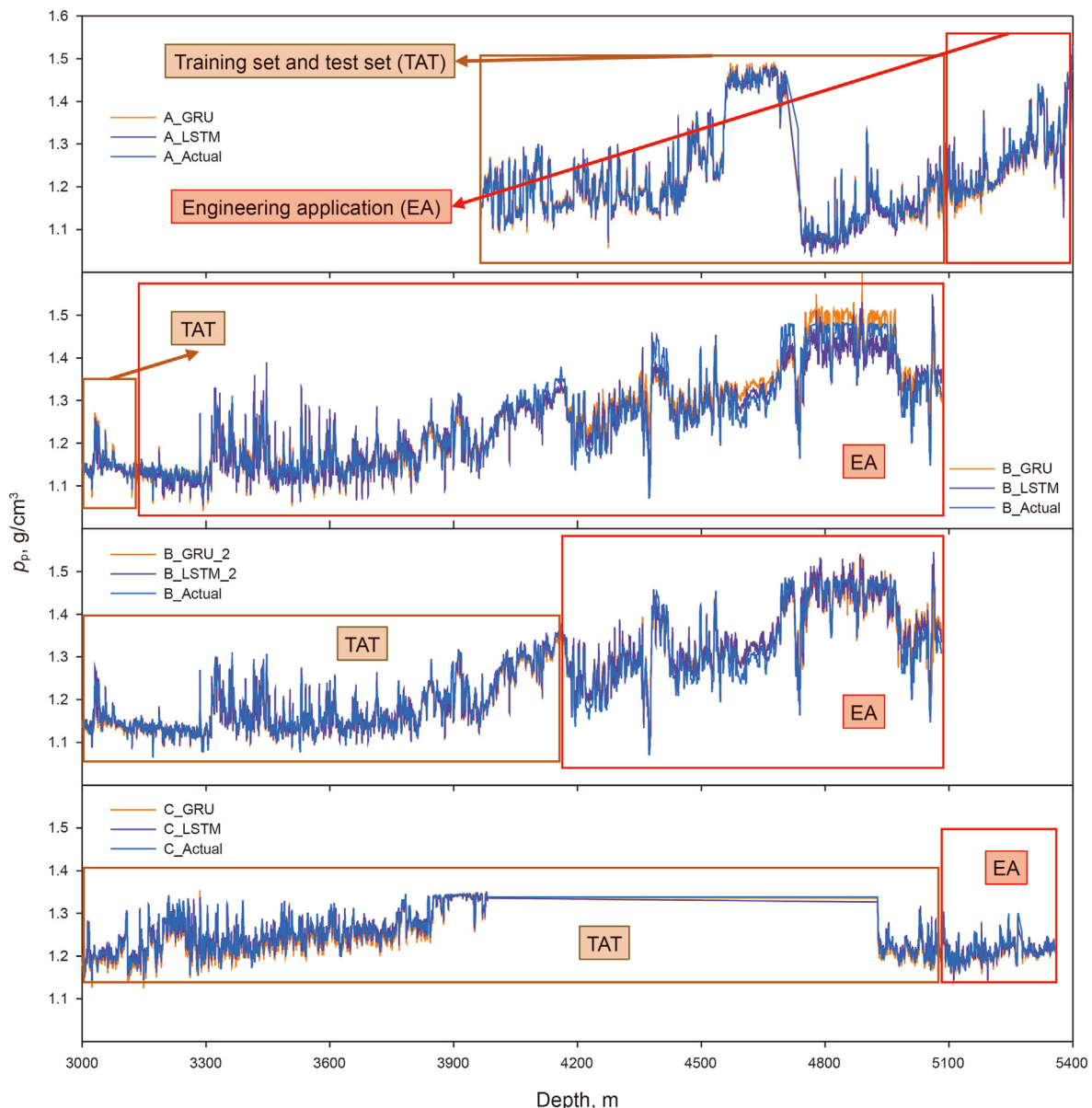


Fig. 9. Comparison between the predicted results of the three wells and the true values.

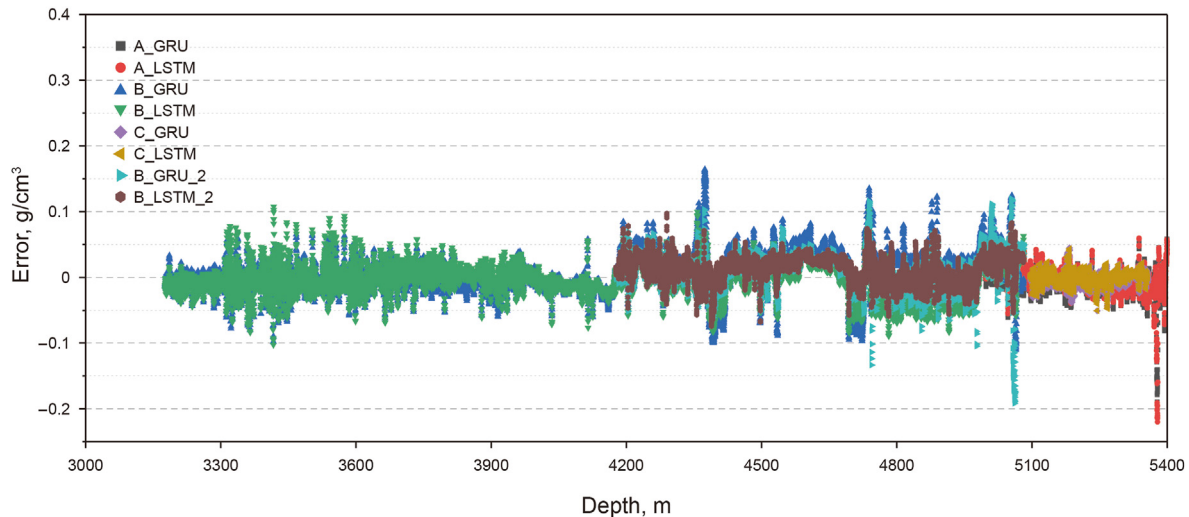


Fig. 10. The error between the predicted results of three wells and the true value.

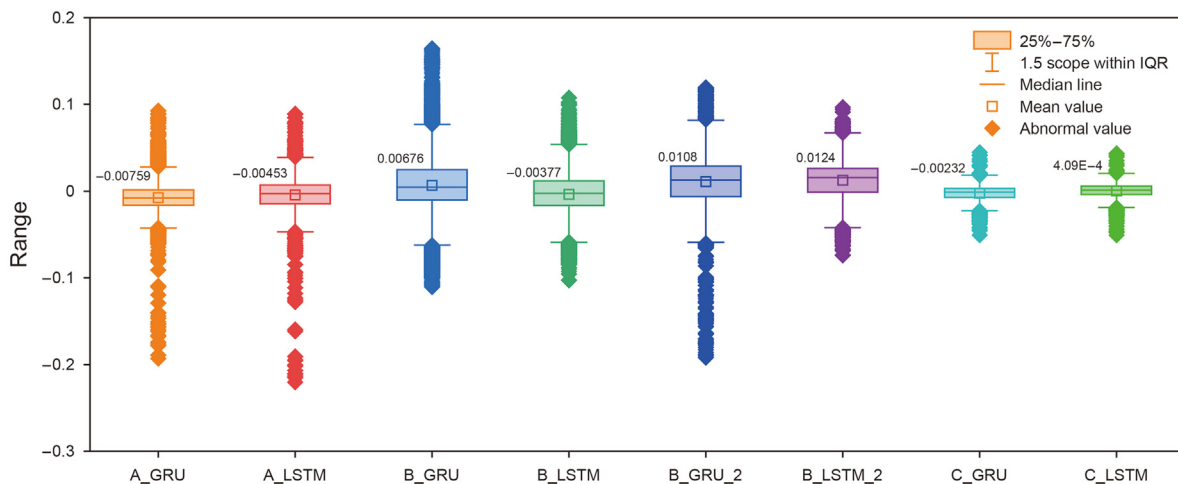


Fig. 11. Distribution of predicted results error.

petrophysical theory, whether based on empirical formulas or the methods proposed in this study, adhere to natural physical principles. It further underscores the feasibility of embedding petrophysical theory into machine learning models.

It is noteworthy that while the predicted trends generally align with post-drilling corrections, some segments in the wells show slight overestimations or underestimations of formation pore pressure values. This discrepancy arises because actual formation pore pressure values for all segments are not directly obtainable. The formation pore pressure values used for training and validation in this study are derived from comprehensive inversion of logging data and drilling complexities. While highly informative for engineering applications, these values do not represent exact formation pore pressure values. For example, in Well B, both models trained on the first 10000 records (B\_GRU, B\_LSTM) and the first 20000 records (B\_GRU\_2, B\_LSTM\_2) predict higher formation pore pressure values between 4423.8 and 4690.3 m compared to corrected formation pore pressure values. Reviewing Well B's drilling logs revealed a kick at 4890 m, resulting in a mud pit increase of 1.8 m<sup>3</sup>, a shut-in pressure of 9.5 MPa, and a static pressure of 4.15 MPa after hard shut-in. After a three-week well control

operation, drilling fluid density was raised from 1.45 to 1.54 g/cm<sup>3</sup>, and conditions returned to normal. During post-drilling formation pore pressure corrections, only the formation pore pressure at the kick point was adjusted, not the entire preceding normal drilling section's formation pore pressure.

B\_GRU exhibited significant errors in predicting formation pore pressure in the lower formations. This is expected as the model, trained on 10000 out of 29032 records from Well B, did not fully capture the principles underlying the prediction of formation pore pressure based on the selected parameters. Specifically, the model failed to learn the relationship between formation pore pressure in deeper formations and the various parameters, leading to substantial errors when predicting formation pore pressure below 4300 m, with errors remaining around 0.1 g/cm<sup>3</sup> between 4700 and 5000 m. It is noteworthy that while this model showed large errors in predicting deeper formations, it accurately predicted formation pore pressure within the lowermost 1000 m beneath the drill bit. In contrast, both B\_GRU\_2 and B\_LSTM\_2 demonstrated good adaptability in predicting high-pressure intervals in the lower formations. This indicates that with sufficient data, the proposed method in this study effectively predicts formation pore pressure in the



**Table 3**  
Performance of various models.

|                | A_GRU          | A_LSTM  | B_GRU   | B_LSTM  | B_GRU_2 | B_LSTM_2 | C_GRU   | C_LSTM          |
|----------------|----------------|---------|---------|---------|---------|----------|---------|-----------------|
| RMSE           | 0.02127        | 0.02227 | 0.03072 | 0.02487 | 0.03051 | 0.02338  | 0.00937 | <b>0.008964</b> |
| MAE            | 0.01415        | 0.01464 | 0.02273 | 0.01914 | 0.02361 | 0.01970  | 0.00676 | <b>0.00656</b>  |
| MAPE           | 1.08635        | 1.12355 | 1.77914 | 1.48314 | 1.79964 | 1.50620  | 0.55633 | <b>0.53982</b>  |
| R <sup>2</sup> | <b>0.94928</b> | 0.94444 | 0.92040 | 0.94784 | 0.89585 | 0.93884  | 0.82165 | 0.83682         |
| AAPD           | 1.10270        | 1.13934 | 1.75709 | 1.49085 | 1.77667 | 1.48305  | 0.55915 | <b>0.54024</b>  |

lower formations beneath the drill bit. These results also suggest that the models continuously update as drilling progresses, gradually converging towards real-world conditions.

The error distribution of predictions from all eight models converges 0.1 g/cm<sup>3</sup> across the three wells, with the majority of data points converging around 0.05 g/cm<sup>3</sup>. Although these models may not precisely invert formation pore pressure for deeper formations with partial data, they can accurately forecast the distribution of formation pore pressure within a certain distance beneath the drill bit. Such models are effective in practical drilling operations for avoiding complex incidents such as kicks and losses.

The models applying the method proposed in this study to three wells were evaluated using five performance metrics to assess their application in detecting formations at remote depths, as summarized in Table 3 and illustrated in Fig. 12 for model performance evaluation. The method proposed in this study demonstrated favorable outcomes across all three wells. For instance, the GRU model applied to Well A performed optimally in four out of eight models considered in this study.

It is important to note that based on the model evaluation results, B\_GRU\_2 and B\_LSTM\_2 appear to exhibit poorer performance compared to B\_GRU and B\_LSTM. However, this observation is somewhat misleading due to the nature of the data. B\_GRU and B\_LSTM were applied within the depth range of 3176.7–5081.4 m, encompassing both normal-pressure intervals and the entire high-pressure interval. The accurate predictions in normal-pressure intervals offset the poorer predictions in the high-pressure interval. In contrast, B\_GRU\_2 and B\_LSTM\_2 were applied within the depth range of 4176.7–5081.4 m, which exclusively covered high-pressure intervals. Moreover, incidents of kicks occurred during drilling, and some high-pressure intervals were not corrected to reflect actual conditions, resulting in approximately 260 m length discrepancy between predicted and corrected formation pore pressure values. Consequently, B\_GRU\_2 and B\_LSTM\_2 appear to exhibit poorer performance compared to B\_GRU and B\_LSTM based on their performance metrics.

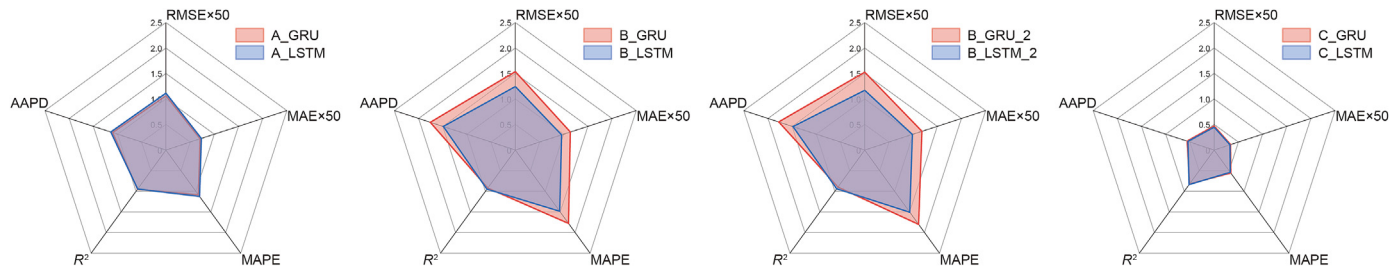
4.3. Explanation of machine learning models

This study employs the SHAP to elucidate the machine learning models B\_GRU, B\_LSTM, B\_GRU\_2, and B\_LSTM\_2 in predicting formation pore pressure across the entire wellbore. Fig. 13 depicts bee swarm plots of SHAP analyses for the four models, illustrating

the contribution of each feature to the accurate prediction of formation pore pressure. Fig. 14 presents dependency plots showing the relationship between each feature and PVSEI for the four models. Here, features are colored from blue (indicating lower feature values) to red (indicating higher feature values).

From the SHAP analysis results of the four models, Depth consistently exhibits significant importance across all models, indicating its stable and strong impact on predicting formation pore pressure. Depth shows a clear positive correlation with model outputs, aligning with geological principles that relate depth to formation pore pressure (Zhao et al., 2023). The GR feature demonstrates a complex relationship in predicting formation pore pressure, as evident from the overlapping colored regions in Fig. 14, showing both positive and negative impacts on the models. DT consistently ranks second in SHAP contributions across all models, highlighting its crucial role in predicting formation pore pressure. DT is proportional to SHAP value, which is consistent with the law that the abnormal increase of DT in nature indicates the existence of under compacted high-pressure fluid. CNCF and Den show relatively lower importance in all models but still contribute to the accurate prediction of formation pore pressure. It is noteworthy that this study incorporates the PVSEI feature, not traditionally used in models, as an input parameter, and it demonstrates significant importance in models B\_GRU, B\_LSTM, B\_GRU\_2 and B\_LSTM\_2. PVSEI exhibits a pronounced positive correlation with formation pore pressure, consistent with trends observed in Fig. 3 of the original data, indicating that machine learning has captured the relationship between PVSEI and formation pore pressure. It shows that PVSEI provides a strong contribution to the accurate prediction of formation pore pressure.

Comparing the SHAP value bee swarm plots of the four models reveals that overall feature importance ranking and their impact on the models are quite similar. Minor differences may arise from variations in model architecture and the size of the training dataset used. The primary principle of the Eaton method for predicting formation pore pressure involves measuring the acoustic travel time through the formation, comparing it with the travel time under normal pressure conditions, and incorporating changes in vertical and effective stresses to determine formation pore pressure. Across all models, Depth and DT consistently occupy the first and second positions in terms of contribution, indicating that the models in this study reflect certain natural geological laws.



**Fig. 12.** Evaluation of model performance (values of RMSE and MAE are multiplied by 50 for better presentation of results).

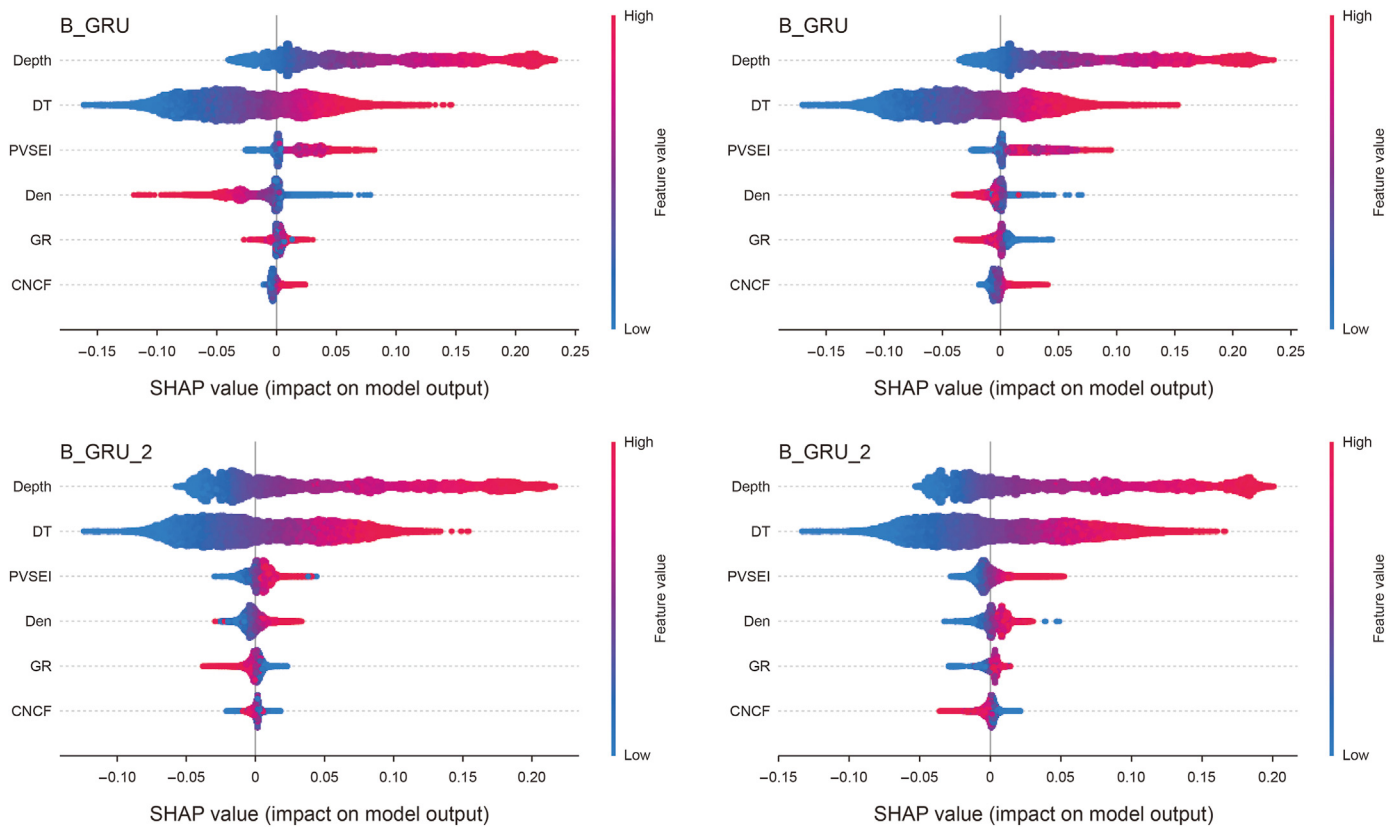


Fig. 13. Bee swarm plots of the four model SHAP analyses.

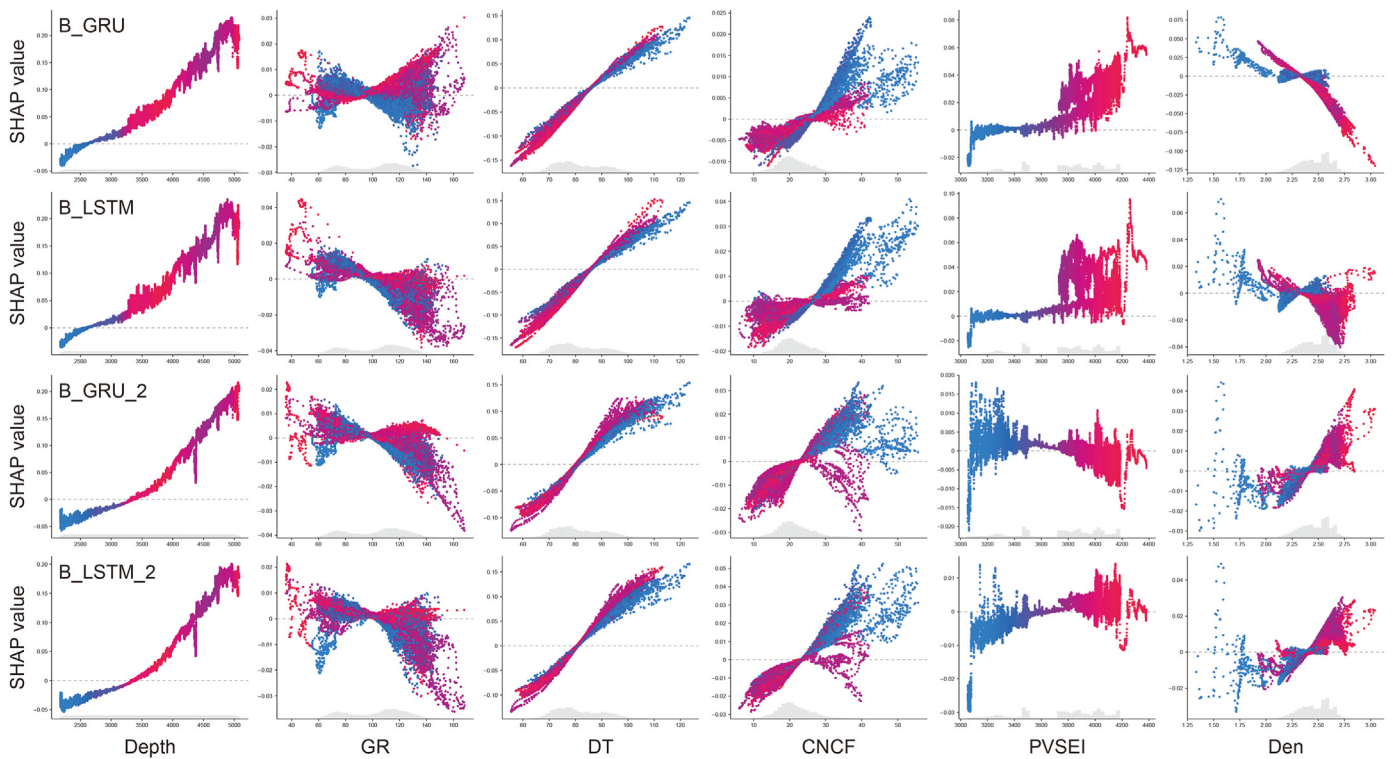


Fig. 14. SHAP dependency graphs for features in each model (the color represents the actual value of PVSEI).

Fig. 15 is the force diagram of four models at 4680 m. At 4680 m, since there is no subsequent overflow high pressure after drilling to correct the predicted Pp at this location. All four models predict the pressure increase identified by the traditional method error at this location. This indicates that the three models consistently capture similar formation pressure characteristics. Among them, the B\_GRU and B\_LSTM models trained based on the first 10000 samples show similar characteristic distribution patterns, and the final predicted values are both 1.32 g/cm<sup>3</sup>. In both models, PVSEI has a large contribution, which indicates that seismic data can provide some geological insights that traditional methods cannot discover, further enriching the understanding and prediction methods of formation pressure. As the number of samples reaches 20000, the Depth feature has a more important contribution. However, DT, Den, and PVSEI are still the top four contributions. This indicates that the coupling of Depth, DT, Den, and PVSEI can better understand the change of pressure.

In summary, applying SHAP analysis in formation pore pressure prediction models provides valuable insights into the relative importance and impacts of input features. These insights can effectively guide model improvement, feature selection, or feature engineering to achieve more accurate predictions of formation pore pressure.

#### 4.4. Discussion and limitations

Despite providing the insights above, future research still faces several limitations that need to be addressed. Considering the inability to directly obtain extensive field measurements of formation pore pressure, the models proposed in this study use supervised learning labels derived from theoretical calculations, combined with log data and corrections based on drilling complexities. While such results effectively reflect the distribution of formation pore pressure in engineering applications, they do not necessarily represent the actual distribution patterns of formation pore pressure. In future research, obtaining extensive field

measurements of formation pore pressure remains a significant challenge for petroleum geologists and engineers. Additionally, the physical theory employed in this study is based on the Eaton method, which is suitable for undercompaction-related overpressure. Directly applying this model to other blocks or regions may yield inaccurate predictions. When using this method to predict pore pressure in different regions, it is essential to first identify the overpressure mechanism specific to the area and then select an appropriate physical theory model to integrate with the machine learning model for Pp prediction. Therefore, understanding the mechanisms behind overpressure formation remains a key focus for future research.

Furthermore, embedding physical theories into the loss function may also raise controversies and issues. Machine learning is data-driven, aiming to fully learn and uncover hidden relationships and patterns behind the data when constructing models. However, in practical engineering operations, theoretical calculation methods often involve empirical formulas derived from extensive operations. For instance, there are many methods to determine normal sonic travel time curves, with some scholars using curves established across entire blocks and others fitting measured sonic travel times in segments sorted in ascending order (Li et al., 2022b). Both methods may deviate from reality. Introducing such empirical formulas into data-driven methods may potentially hinder the thorough exploration of relationships and patterns hidden within the data by machine learning.

Although this study employs SHAP methods to explain formation pore pressure prediction models, there are still several limitations. Firstly, there are computational complexity and time expenses. In this study, it took 70 min to perform SHAP analysis on all data from Well B using a single model. Extensive computations consume considerable time, and with an increase in feature parameters, this time will exponentially grow. Additionally, SHAP analysis depends on model complexity; the nonlinearity and complexity of neural network models complicate interactions between features. SHAP values assume independence between



Fig. 15. Force diagrams of four models at 4680 m.

features or calculate through conditional independence, which may not accurately reflect complex feature interactions in actual models. Moreover, the interpretability of SHAP methods is limited. SHAP values provide feature contribution scores to model outputs, which may still be challenging for non-technical users to directly comprehend. Despite attempts by SHAP to simplify the explanation process, interpreting complex neural network models may remain overly abstract.

## 5. Conclusions

This study proposes a novel methodology for predicting formation pore pressure ahead of the drill bit by integrating petrophysical theory into machine learning models driven by seismic and LWD data. The developed rock physics-informed model effectively bridges machine learning with geological principles, ensuring predictions adhere to physical laws by embedding prior geological knowledge through a refined loss function. Operating within the downhole environment, the model processes real-time sensor data directly and transmits only predicted pore pressure values to the surface, addressing significant constraints in downhole data transmission bandwidth.

A detailed examination of drilling history and pore pressure distribution reveals that traditional methods, such as the Eaton method, are limited in their ability to capture complex formation pressures accurately. PVSEI is demonstrated to be a reliable indicator of pore pressure fluctuations, effectively capturing trends not reflected in traditional parameters such as GR and DT. Tests using 65,837 data records from three Bohai Oilfield wells confirm the effectiveness of this approach: the GRU and LSTM models exhibit strong adaptability across varying depths. Nearly all model prediction errors converge around 0.05 g/cm<sup>3</sup>, with the maximum root mean square error (RMSE) being 0.03072 and the minimum RMSE being 0.008964. Moreover, predictive accuracy improves as drilling progresses and the volume of available data increases.

SHAP analysis confirms that key features such as Depth and DT exert significant influence on model outputs, consistent with Eaton's foundational principles of vertical and effective stress variation. Additionally, PVSEI proves essential for pore pressure prediction, highlighting its potential for enhancing predictive accuracy and broadening the scope of parameter selection in future formation pressure modeling endeavors. Analyses based on different data samples further indicate that combining Depth, DT, Den, and PVSEI can provide a clearer understanding of pressure variations. This study demonstrates the feasibility of embedding petrophysical theory into machine learning frameworks, offering a robust pathway for more precise, real-time formation pore pressure prediction in complex drilling environments.

## CRediT authorship contribution statement

**Xu-Yue Chen:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Cheng-Kai Weng:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis. **Lin Tao:** Resources, Data curation. **Jin Yang:** Supervision, Conceptualization. **De-Li Gao:** Supervision, Conceptualization. **Jun Li:** Supervision, Conceptualization.

## Data availability

The data that has been used is confidential.

## Conflict of interest

We hereby confirm that this manuscript is our original work and none of the material in the paper has been published or is under consideration for publication elsewhere, and there are no interest conflicts. Any details of funding agencies have already stated clearly in the manuscript.

## Acknowledgments

This work was financially supported by the National Natural Science Foundation of China (Grant numbers: 52174012; 52394250; 52394255; 52234002; U22B20126; 51804322).

## References

- Ahmed, S.A., Mahmoud, A.A., Elkhatny, S., et al., 2019. Prediction of Pore and Fracture Pressures Using Support Vector Machine. International Petroleum Technology Conference, Beijing, China. <https://doi.org/10.2523/IPTC-19523-MS>.
- Ahmed, A.A., Salaheldin, E.S., Abdulazeez, A.A., 2021. Formation pressure prediction from mechanical and hydraulic drilling data using artificial neural networks. In: ARMA US Rock Mechanics/Geomechanics Symposium. Muscat, Oman. <https://doi.org/10.2118/191391-18IHFT-MS>.
- Bialek, J., Bujalski, W., Wojdan, K., et al., 2022. Dataset level explanation of heat demand forecasting ANN with SHAP. Energy 261, 125075. <https://doi.org/10.1016/j.energy.2022.125075>.
- Biot, M.A., 1941. General theory of three-dimensional consolidation. J. Appl. Phys. 12 (2), 155–164. <https://doi.org/10.1063/1.1712886>.
- Bommidi, B.S., Teeparthi, K., Kosana, V., 2023. Hybrid wind speed forecasting using ICEEMDAN and transformer model with novel loss function. Energy 265, 126383. <https://doi.org/10.1016/j.energy.2022.126383>.
- Bowers, G.L., 1995. Pore pressure estimation from velocity data: accounting for overpressure mechanisms besides undercompaction. SPE Drilling 10 (2), 89–95. <https://doi.org/10.2118/27488-PA>.
- Cao, S., Wang, C., Niu, Q., et al., 2024. Enhancing pore pressure prediction accuracy: a knowledge-driven approach with temporal fusion transformer. Geoenergy Sci. Eng. 238, 212839. <https://doi.org/10.1016/j.geoen.2024.212839>.
- Chen, X., Weng, C., Du, X., et al., 2023. Prediction of the rate of penetration in offshore large-scale cluster extended reach wells drilling based on machine learning and big-data techniques. Ocean Eng. 285, 115404. <https://doi.org/10.1016/j.oceaneng.2023.115404>.
- Chen, X., Du, X., Weng, C., et al., 2024. A real-time drilling parameters optimization method for offshore large-scale cluster extended reach drilling based on intelligent optimization algorithm and machine learning. Ocean Eng. 291, 116375. <https://doi.org/10.1016/j.oceaneng.2023.116375>.
- Das, B., Chatterjee, R., 2018. Mapping of pore pressure, in-situ stress and brittleness in unconventional shale reservoir of Krishna-Godavari basin. J. Nat. Gas Sci. Eng. 50, 74–89. <https://doi.org/10.1016/j.jngse.2017.10.021>.
- Deng, S., Pan, H., Wang, H., et al., 2024. A hybrid machine learning optimization algorithm for multivariable pore pressure prediction. Pet. Sci. 21 (1), 535–550. <https://doi.org/10.1016/j.petsci.2023.09.001>.
- Eaton, B.A., 1972. The effect of overburden stress on geopressure prediction from well logs. J. Petrol. Technol. 24 (8), 929–934. <https://doi.org/10.2118/3719-PA>.
- Eaton, B.A., 1975. The Equation for Geopressure Prediction from Well Logs. Fall Meeting of the Society of Petroleum Engineers of AIME, Dallas, Texas. <https://doi.org/10.2118/5544-MS>.
- Gholilou, A., Vialle, S., Madadi, M., 2017. Determination of safe mud window considering time-dependent variations of temperature and pore pressure: analytical and numerical approaches. J. Rock Mech. Geotech. Eng. 9 (5), 900–911. <https://doi.org/10.1016/j.jrmge.2017.02.002>.
- Guo, Q.B., Hao, X.B., Wu, C., et al., 2020. Deep EM method for proactively prediction of resistivity ahead of bit to determine salt bottom position. In: International Petroleum Technology Conference, Dhahran, Kingdom of Saudi Arabia. <https://doi.org/10.2523/IPTC-19616-Abstract>.
- Guo, J., Li, M., Zhuang, M., et al., 2023. Rock physics model for velocity–pressure relations and its application to shale pore pressure estimation. Petrol. Explor. Dev. 50 (2), 404–418. [https://doi.org/10.1016/S1876-3804\(23\)60396-9](https://doi.org/10.1016/S1876-3804(23)60396-9).
- Hagiwara, T., 2018. Detection sensitivity and new concept of deep reading look-ahead look-around geosteering tool. In: SEG International Exposition and Annual Meeting, Anaheim, California. <https://doi.org/10.1190/segam2018-2981412.1>.
- He, Y., He, Z., Tang, Y., et al., 2023. Shale gas production evaluation framework based on data-driven models. Pet. Sci. 20 (3), 1659–1675. <https://doi.org/10.1016/j.petsci.2022.12.003>.
- Hottmann, C., Johnson, R., 1965. Estimation of formation pressures from log-derived shale properties. J. Petrol. Technol. 17 (6), 717–722. <https://doi.org/10.2118/1110-PA>.
- Hubbert, M., Rubey, W.W., 1959. Role of fluid pressure in mechanics of overthrust faulting: I. Mechanics of fluid-filled porous solids and its application to



- overthrust faulting. *GSA Bulletin* 70 (2), 115–166. [https://doi.org/10.1130/0016-7606\(1959\)70\[115:ROFPIM\]2.0.CO;2](https://doi.org/10.1130/0016-7606(1959)70[115:ROFPIM]2.0.CO;2).
- Jorden, J., Shirley, O., 1966. Application of drilling performance data to overpressure detection. *J. Petrol. Technol.* 18 (11), 1387–1394. <https://doi.org/10.2118/1407-PA>.
- Kannangara, K.P.M., Zhou, W., Ding, Z., et al., 2022. Investigation of feature contribution to shield tunneling-induced settlement using Shapley additive explanations method. *J. Rock Mech. Geotech. Eng.* 14 (4), 1052–1063. <https://doi.org/10.1016/j.jrmge.2022.01.002>.
- Khalil, H., Seydoux, J., Denichou, J., et al., 2018. Successful implementation of real-time look-ahead resistivity measurements in the North Sea. In: *SPE Norway Subsurface Conference*, Bergen, Norway. <https://doi.org/10.2118/191340-MS>.
- Lan, P., Su, J., Zhang, S., 2023. Surrogate modeling for unsaturated infiltration via the physics and equality-constrained artificial neural networks. *J. Rock Mech. Geotech. Eng.* 16 (6), 2282–2295. <https://doi.org/10.1016/j.jrmge.2023.09.014>.
- Li, D., Ma, J., Li, L., et al., 2022a. Prediction method of formation pressure in abnormal pressure reservoir in Bozhong sag. *Prog. Geophys.* 37 (3), 1266–1273. <https://doi.org/10.6038/pg2022FF0043>.
- Li, G., Song, X., Tian, S., et al., 2022b. Intelligent drilling and completion: a review. *Engineering* 18, 33–48. <https://doi.org/10.1016/j.eng.2022.07.014>.
- Li, Q., Fu, J., Peng, C., et al., 2023. A deep learning approach for abnormal pore pressure prediction based on multivariate time series of kick. *Geoenery Sci. Eng.* 226, 211715. <https://doi.org/10.1016/j.geoen.2023.211715>.
- Liu, H., Yuan, F., Jiang, Y., et al., 2019. Mechanisms for overpressure generated by the undercompaction of paleogene strata in the Baxian depression of Bohai Bay Basin, China. *Mar. Petrol. Geol.* 99, 337–346. <https://doi.org/10.1016/j.marpetgeo.2018.10.001>.
- Lockhart, L.P., Flemings, P.B., Nikolinakou, M., et al., 2023. Velocity-based pore pressure prediction in a basin with late-stage erosion: Delaware Basin, US. *Mar. Petrol. Geol.* 150, 106159. <https://doi.org/10.1016/j.marpetgeo.2023.106159>.
- Loutfi, A.A., Sun, M., Loutfi, I., et al., 2022. Empirical study of day-ahead electricity spot-price forecasting: insights into a novel loss function for training neural networks. *Appl. Energy* 319, 119182. <https://doi.org/10.1016/j.apenergy.2022.119182>.
- Ma, T., Xiang, G., Gui, J., et al., 2024. Physics-constrained distributed neural network model for 3D in-situ stress prediction. *Chin. J. Geophys.* 67 (8), 3211–3228. <https://doi.org/10.6038/cjg2023R0132> (in Chinese).
- Mahmoud, A.A., Alzayer, B.M., Panagopoulos, G., et al., 2024. A new empirical correlation for pore pressure prediction based on artificial neural networks applied to a real case study. *Processes* 12 (4), 664. <https://doi.org/10.3390/pr12040664>.
- Pu, X., Zhao, X., Wang, J., et al., 2020. Reservoirs properties of slump-type sub-lacustrine fans and their main control factors in first member of Paleogene Shahejie Formation in Binhai area, Bohai Bay Basin, China. *Petrol. Explor. Dev.* 47 (5), 977–989. [https://doi.org/10.1016/S1876-3804\(20\)60110-0](https://doi.org/10.1016/S1876-3804(20)60110-0).
- Ribeiro, K.S., Vieira, M.B., Villela, S.M., et al., 2024. Composite loss function for 3-D poststack seismic data compression. *Comput. Geosci.* 189, 105620. <https://doi.org/10.1016/j.cageo.2024.105620>.
- Song, Z., Cao, S., Yang, H., 2024. An interpretable framework for modeling global solar radiation using tree-based ensemble machine learning and Shapley additive explanations methods. *Appl. Energy* 364, 123238. <https://doi.org/10.1016/j.apenergy.2024.123238>.
- Sun, B., Cheng, Y., Han, Z., et al., 2024. Experimental study of a multimechanism formation pressure prediction model influenced by shale content. *Geoenery Sci. Eng.* 233, 212486. <https://doi.org/10.1016/j.geoen.2023.212486>.
- Terzaghi, K., 1943. *Theoretical Soil Mechanics*. John Wiley & Sons. <https://doi.org/10.1002/9780470172766>.
- Wang, Z., Wang, R., 2015. Pore pressure prediction using geophysical methods in carbonate reservoirs: current status, challenges and way ahead. *J. Nat. Gas Sci. Eng.* 27, 986–993. <https://doi.org/10.1016/j.jngse.2015.09.032>.
- Wang, Y., Ye, H., 2024. Current status and development trend of measurement & control while drilling technology. *Petrol Drill Technol* 52 (1), 122–129. <https://doi.org/10.11911/syztjs.2024017> (in Chinese).
- Wang, L., Deng, S., Xie, G., et al., 2022. A new deep-reading look-ahead method in electromagnetic logging-while-drilling using the scattered electric field from magnetic dipole antennas. *Pet. Sci.* 19 (1), 180–188. <https://doi.org/10.1016/j.petsci.2021.09.035>.
- Wang, Z., Chen, G., Zhang, R., et al., 2023. Early monitoring of gas kick in deepwater drilling based on ensemble learning method: a case study at South China Sea. *Process Saf Environ* 169, 504–514. <https://doi.org/10.1016/j.psep.2022.11.024>.
- Xiong, Z., Yao, J., Huang, Y., et al., 2024. A wind speed forecasting method based on EMD-MGM with switching QR loss function and novel subsequence superposition. *Appl. Energy* 353, 122248. <https://doi.org/10.1016/j.apenergy.2023.122248>.
- Xu, Y., Yang, L., Xu, J., et al., 2024. Prediction method for formation pore pressure based on transfer learning. *Geoenery Sci. Eng.* 236, 212747. <https://doi.org/10.1016/j.geoen.2024.212747>.
- Xue, Y., Wang, D., 2020. Formation conditions and exploration direction of large natural gas reservoirs in the oil-prone Bohai Bay Basin, East China. *Petrol. Explor. Dev.* 47 (2), 280–291. [https://doi.org/10.1016/S1876-3804\(20\)60046-5](https://doi.org/10.1016/S1876-3804(20)60046-5).
- Zhang, J., 2011. Pore pressure prediction from well logs: methods, modifications, and new approaches. *Earth Sci. Rev.* 108 (1), 50–63. <https://doi.org/10.1016/j.earscirev.2011.06.001>.
- Zhang, C., Zhang, R., Zhu, Z., et al., 2023. Bottom hole pressure prediction based on hybrid neural networks and Bayesian optimization. *Pet. Sci.* 20 (6), 3712–3722. <https://doi.org/10.1016/j.petsci.2023.07.009>.
- Zhang, X., Lu, Y., Jin, Y., et al., 2024. An adaptive physics-informed deep learning method for pore pressure prediction using seismic data. *Pet. Sci.* 21 (2), 885–902. <https://doi.org/10.1016/j.petsci.2023.11.006>.
- Zhao, X., Chen, X., Lan, Z., et al., 2023. Pore pressure prediction assisted by machine learning models combined with interpretations: a case study of an HTHP gas field, Yinggehai Basin. *Geoenery Sci. Eng.* 229, 212114. <https://doi.org/10.1016/j.geoen.2023.212114>.
- Zhou, K., Zhang, J., Ren, Y., et al., 2020. A gradient boosting decision tree algorithm combining synthetic minority oversampling technique for lithology identification. *Geophysics* 85 (4), 147–158. <https://doi.org/10.1190/geo2019-0429.1>.