

Article

A Collapse Pressure Prediction Method Based on Virtual-Well Constraints and Deep Sequence Learning

Ning Li ^{1,2,3,4}, Jiaqi Luo ⁵, Wentong Fan ^{1,2,3,4}, Yang Xia ^{5,*} and Zhenyu Zhang ⁵

¹ PetroChina Tarim Oilfield Company, Korla 841000, China; lining-tlm@petrochina.com.cn (N.L.); fanwt-tlm@petrochina.com.cn (W.F.)

² R&D Center for Ultra Deep Complex Reservoir Exploration and Development, CNPC, Korla 841000, China

³ Engineering Research Center for Ultra-Deep Complex Reservoir Exploration and Development, Korla 841000, China

⁴ Xinjiang Key Laboratory of Ultra-Deep Oil and Gas, Korla 841000, China

⁵ College of Petroleum Engineering, China University of Petroleum (Beijing), Beijing 102249, China; luojiaqi000927@163.com (J.L.); 13261441199@163.com (Z.Z.)

* Correspondence: xiyang@vip.126.com

Abstract

The collapse pressure equivalent density is a key parameter for determining the safe drilling fluid density window and evaluating wellbore stability. To address the limited number of drilled wells, the lack of continuous geomechanical labels, and the complex relationship between seismic responses and collapse pressure, this study proposes a collapse pressure prediction method based on virtual-well constraints and deep sequence learning. Virtual-well samples were constructed from the acoustic impedance distributions of drilled wells, synthetic seismic records were generated through actual seismic wavelet extraction and reflection-coefficient convolution, and collapse-pressure equivalent-density labels were calculated using a rock-mechanics model, thereby forming a large-scale training dataset. Random Forest, Polynomial Regression, Support Vector Regression, and CNN-MultiLSTM models were compared. The CNN-MultiLSTM framework was further optimized in terms of input range, feature channels, and sequence structure, resulting in a whole-well sequence model combining multichannel inputs with residual dilated convolutions. The optimized model achieved an MAE of 0.00262 g/cm³, an RMSE of 0.00352 g/cm³, a MAPE of 0.241%, and an R² of 0.9774 on the validation set. In independent validation using an actual well not involved in training, the model achieved an MAE of 0.0182 g/cm³, an RMSE of 0.0278 g/cm³, a MAPE of 1.640%, and an R² of 0.7356 within the primary target interval of 7900.000–7994.100 m. The predicted profile reproduced the main depth-dependent variation of the calculated collapse-pressure equivalent density, although larger deviations occurred in locally abrupt intervals. Analysis outside the primary target interval further showed that the model could respond to high-collapse-pressure anomalies. Overall, integrating virtual-well constraints, rock-mechanics-based labeling, and whole-well sequence learning provides a feasible approach for collapse-pressure prediction in undrilled areas and drilling fluid density design, while further multi-well validation is required to assess cross-well generalization.

Keywords: collapse pressure; virtual well; synthetic seismic record; deep sequence learning; whole-well sequence model



Academic Editor: Angelo Luongo

Received: 12 June 2026

Revised: 30 July 2026

Accepted: 2 August 2026

Published: 4 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

1. Introduction

Collapse pressure is one of the key parameters for determining the safe drilling fluid density window and evaluating wellbore stability. Previous studies have investigated wellbore failure mechanisms from the perspectives of in-situ stress, borehole inclination, rock strength, and wellbore geometry [1–4]. Subsequent studies further incorporated advanced failure criteria and formation anisotropy into wellbore stability analysis [5,6]. Therefore, accurate prediction of the collapse pressure and its equivalent mud density in the target interval is not only the basis for wellbore stability evaluation, but also an important prerequisite for optimizing drilling fluid density, reducing downhole complications, and improving drilling efficiency.

Traditional methods for collapse pressure prediction are generally based on poroelastic mechanics and rock strength criteria. These methods have clear rock-mechanics significance; however, in practical applications, the required parameters are often difficult to obtain continuously in undrilled areas, making such models susceptible to parameter uncertainty. To overcome these limitations, researchers have proposed empirical methods for estimating rock strength and pore pressure from data such as acoustic velocity [7–10]. For example, a certain statistical relationship exists between the strength of sedimentary rocks and petrophysical logging properties, but this relationship is affected by lithology, diagenesis, compaction degree, and fluid pressure mechanisms, and therefore needs to be calibrated using data from the study area [7,8]. In pore pressure prediction, the relationship between velocity and effective stress can reflect undercompaction or abnormal fluid pressure; however, the velocity responses are not entirely consistent under different overpressure mechanisms [9,10]. Therefore, although traditional prediction methods can provide physical interpretations, they are insufficient on their own to address problems such as limited parameter availability, strong regional heterogeneity, and unstable lateral prediction accuracy in complex formations.

With the continuous accumulation of geophysical, logging, and drilling engineering data, artificial intelligence methods have gradually become important tools for predicting complex geoenvironmental parameters. Machine learning can extract hidden patterns from high-dimensional, nonlinear, and noisy data, and has shown strong applicability in data analysis for solid Earth science and petroleum engineering [11–13]. Meanwhile, studies on theory-guided data science have emphasized that prior knowledge should be incorporated into the model training process to reduce the uncertainty of purely black-box models and improve their generalization capability [14,15]. In terms of specific models, Random Forest reduces the variance of individual decision trees through bootstrap sampling and multi-tree ensemble learning, making it suitable for nonlinear regression and multivariable coupling problems [16]. Support Vector Machine is based on structural risk minimization and can establish nonlinear mappings through kernel functions; accordingly, Support Vector Regression exhibits strong generalization capability in the prediction of continuous variables with small samples [17,18]. These models provide optional baseline methods for collapse pressure prediction. However, when the input data exhibit vertical sequence correlation, traditional models often struggle to fully represent such relationships.

Convolutional neural networks and recurrent neural networks provide more suitable deep learning architectures for modeling seismic and logging sequences. Convolutional neural networks can extract stable local patterns through local receptive fields and weight sharing, and have been widely applied in image recognition, signal processing, and time-series analysis [19–22]. Particularly for one-dimensional signals, 1D-CNN can automatically extract local waveform features from amplitude sequences, spectral sequences, or logging curves, reducing the dependence of prediction performance on manually engineered features [23]. Long Short-Term Memory networks have been widely applied to nonlinear

sequence modeling and time-series prediction because of their ability to represent long-range dependencies [24–26]. Both seismic data and collapse pressure exhibit continuous sequential characteristics. Therefore, combining the local feature extraction capability of CNN with the sequence memory capability of LSTM can better characterize the non-linear mapping relationship between synthetic seismic responses and collapse pressure equivalent density.

The performance of artificial intelligence-based prediction methods depends not only on the model type, but also on sample construction, data preprocessing, training–validation splitting, and hyperparameter optimization strategies. Deep networks are prone to overfitting when the sample size is limited or the data are highly noisy. Dropout can improve model robustness by randomly masking neuron connections [27]. Hyperparameter optimization methods such as random search have shown that settings including the learning rate, network width, batch size, and regularization parameters can significantly affect model training results [28]. Batch normalization can alleviate internal covariate shift during deep network training and improve convergence stability [29]. Therefore, in collapse pressure prediction, it is necessary to systematically compare Random Forest, Polynomial Regression, Support Vector Regression, and deep sequence models under a unified dataset, unified evaluation metrics, and a reasonable validation strategy, and to further optimize the selected model.

In summary, existing studies have provided an important foundation for collapse pressure prediction from the perspectives of wellbore stability theory, seismic data interpretation, and machine learning modeling. In practical engineering applications, undrilled areas often lack continuous logging and core-test constraints. Key parameters such as in-situ stress, rock strength, and pore pressure must therefore be estimated through empirical relationships or indirect data, causing traditional wellbore stability calculations to be affected by parameter uncertainty [1–10]. Meanwhile, although seismic data have the advantage of laterally continuous coverage, seismic responses are jointly controlled by lithology, pressure, acoustic impedance variations, spectral characteristics, wavelet morphology, and well–seismic calibration [30]. As a result, there is no simple one-to-one correspondence between seismic amplitude and collapse pressure [31,32]. Geng and Wang [33] used multiple seismic attributes and incorporated geomechanical relationships into model training as physical regularization terms to predict wellbore-instability risk levels. In contrast, the present study incorporates geological and geomechanical knowledge primarily at the data-construction stage through impedance-constrained virtual wells, forward-modeled synthetic seismic responses, and rock-mechanics-derived continuous collapse-pressure labels. Thus, the present work focuses on alleviating the shortage of labeled samples and predicting continuous whole-well collapse-pressure equivalent-density profiles rather than directly predicting instability risk from existing seismic-attribute samples. For target intervals with insufficient samples, how to construct large-scale training data with physical consistency remains an issue that needs to be addressed. The generated data should contain seismic response information that reflects lateral variations in the target interval, while also providing reliable mechanical labels and a sufficient sample size. Therefore, constructing large-scale training samples that integrate geological plausibility, realistic seismic responses, and consistency with mechanical constraints is the key to intelligent collapse pressure prediction in undrilled areas.

To address the above issues, this study establishes a prediction workflow based on virtual well data and deep learning models for rapid collapse pressure prediction in the target interval of the study area. First, a large number of virtual wells are generated based on the distribution characteristics of acoustic impedance data from drilled wells in the study area, so that the virtual wells statistically inherit the petrophysical variation characteristics

of the actual formation. Second, seismic wavelets and spectral information are extracted from actual seismic data. Based on reflection coefficients and convolution theory, synthetic seismic data of virtual wells are generated, and collapse pressure labels at corresponding depth points are calculated using a theoretical model of formation collapse pressure. Third, the generated data are subjected to target-interval screening, outlier processing, normalization, and training–validation set splitting. The preprocessed samples are then input into Random Forest, Polynomial Regression, Support Vector Regression, and CNN-MultiLSTM models, and the predictive performance of different models is compared using metrics such as the coefficient of determination. Finally, based on the CNN-MultiLSTM model selected in the comparative experiments, structural optimization is progressively conducted in terms of sample size, local window length, input feature channels, and sequence-modeling range. The prediction strategy is subsequently extended from local-window point-wise prediction to whole-well sequence prediction, and residual dilated convolutions are introduced to enhance long-range depth-context representation. Hyperparameter optimization is then performed for the resulting whole-well sequence model before actual-well validation. This study aims to improve the geological plausibility of virtual well sample generation, enhance the physical interpretability between synthetic seismic data and collapse pressure labels, and provide a methodological reference that integrates data-driven modeling with theoretical constraints for wellbore stability evaluation and drilling fluid density design in undrilled areas.

The main contribution of this study lies in integrating geostatistically constrained virtual-well generation, physics-based collapse-pressure labeling, and deep sequence learning within a unified prediction framework. The proposed workflow further extends local point-wise prediction to whole-well sequence modeling and is validated using an independent actual well. The remainder of this paper is organized as follows: Section 2 describes the data and sample-construction workflow; Section 3 introduces the candidate machine-learning models and evaluation metrics; Section 4 presents model selection and optimization; Section 5 reports the external validation and discusses engineering applicability and limitations; and Section 6 summarizes the main conclusions.

2. Data and Methods

2.1. Data Processing Workflow

The performance of supervised deep learning prediction models often depends on both model complexity and the amount of data available for training. In petroleum exploration and development, the number of wells is usually limited, which restricts the amount of labeled data available for deep learning model training. To overcome this limitation, geophysical prior knowledge must be used to augment labeled data at well locations, thereby generating synthetic labeled data for training deep learning models. This approach is based on the fact that, within a given study area, the physical properties of formation rocks often exhibit a certain degree of similarity and lateral continuity.

This study makes the following two assumptions: ① within a given study area, formation impedance data from different well locations follow a certain probability distribution; ② the impedance data distribution extracted from drilled wells can represent the impedance data distribution within the study area. Based on these assumptions, a large amount of well data can be generated for deep learning training. In this study, the 7900.000–7994.100 m interval of a certain block is selected as the research object. After cleaning the logging data, basic attributes including acoustic slowness, natural gamma ray, bulk density, compressional wave velocity, and acoustic impedance are retained, and the final prediction label is uniformly defined as collapse pressure. The workflow for virtual well data generation is shown in Figure 1.

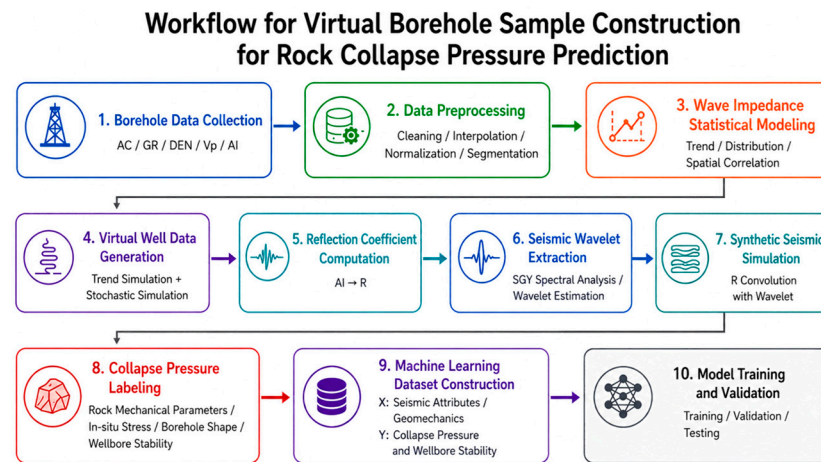


Figure 1. Workflow for virtual well data generation.

2.2. Virtual-Well Acoustic Impedance

Acoustic impedance is an important bridge linking logging data and seismic responses. It is defined as the product of formation density and compressional wave velocity. Let the bulk density and compressional wave velocity at the i th depth sampling point be ρ_i and $v_{p,i}$, respectively. The acoustic impedance can then be expressed as:

$$AI_i = \rho_i \cdot v_{p,i} \quad (1)$$

Therefore, a MATLAB, Version R2022b [34] program was developed to construct virtual-well acoustic impedance using a “distribution pattern + correlated perturbation” strategy. Specifically, the acoustic impedance distribution pattern of drilled wells was first obtained, and then random perturbations with vertical correlation were added in the logarithmic acoustic impedance domain. In this way, the generated curves preserve the overall trend of the target interval while retaining differences among individual wells.

Figure 2 shows the acoustic impedance data of the drilled wells and virtual wells. It can be observed that the two datasets exhibit consistent statistical patterns, indicating that the augmented data are not generated randomly, but are constructed under the constraints of the statistical characteristics of drilled wells. This step ensures that the subsequent reflection coefficients and synthetic seismic responses have a reasonable lithological basis, thereby preventing the model from learning spurious distributions unrelated to actual well data.

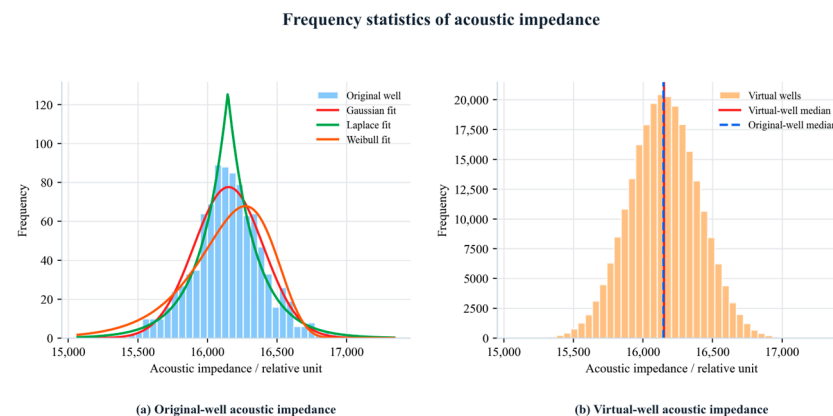


Figure 2. Frequency statistics of acoustic impedance distributions for drilled wells and virtual wells.

2.3. Wavelet Extraction

Wavelet data were extracted based on the drilled-well data, as shown in Figure 3a. Recent studies have emphasized that seismic wavelet extraction should account for depth-dependent characteristics and impedance trends [35]. In this study, the blue solid line in Figure 3a represents the spectral information extracted from the seismic record using the fast Fourier transform [36], while the red solid line represents the fitted wavelet spectrum. The dominant frequency of the wavelet is 14.89 Hz. The extracted wavelet spectrum was then transformed into the time domain using the inverse Fourier transform to obtain the wavelet data, as shown in Figure 3b.

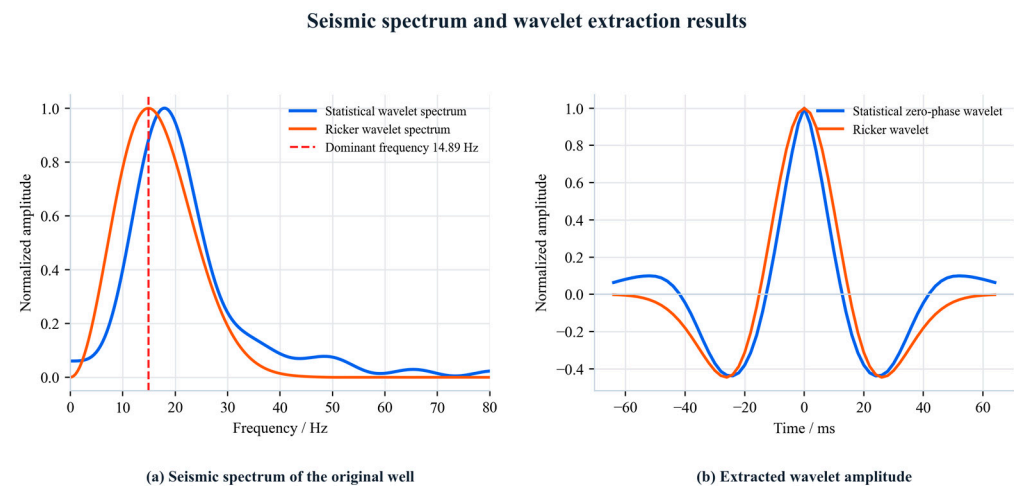


Figure 3. Spectral information of drilled wells and wavelet extraction results.

2.4. Synthetic Seismic Records and Collapse Pressure

After the virtual-well acoustic impedance is obtained, the reflection coefficient can be calculated from adjacent depth sampling points. Let the acoustic impedances of two adjacent points be AI_i and AI_{i+1} , respectively. The reflection coefficient can be expressed as:

$$R_I = \frac{AI_{i+1} - AI_i}{AI_{i+1} + AI_i} \quad (2)$$

The reflection coefficient reflects changes in acoustic properties at formation interfaces and serves as the direct input for generating synthetic seismic responses. The synthetic seismic record of each virtual well can be obtained through time-domain convolution between the wavelet and the reflection coefficient sequence. Let $w(t)$ be the extracted wavelet and $R(t)$ be the reflection coefficient time series. The synthetic seismic record $s(t)$ can then be expressed as:

$$s(t) = w(t) \cdot R(t) \quad (3)$$

To avoid time–depth relationship deviations caused by direct convolution at depth sampling points, the depth-domain reflection coefficients are first converted into the two-way travel time domain according to the logging velocity relationship. They are then convolved with the wavelet under uniform time sampling and subsequently interpolated back to the target depth grid. This procedure ensures that the synthetic seismic records are physically closer to actual seismic responses.

Based on the extracted wavelet of the study area and the reflection coefficients of the virtual wells, the corresponding synthetic seismic records can be obtained through convolution. Subsequently, using the acoustic impedance data of the virtual wells, the basic rock mechanical parameters are calculated by Gardner’s equation [37] and conventional

rock mechanics formulas. On this basis, after obtaining the in-situ stress distribution of the block, the collapse pressure can be determined using the following equation:

$$P_c = \frac{\eta(3\sigma_H - \sigma_h) - 2CK + \alpha P_p(K^2 - 1)}{K^2 + \eta} \quad (4)$$

where P_c denotes the collapse pressure required to maintain wellbore stability; σ_H and σ_h represent the maximum and minimum horizontal in-situ stresses, respectively; C is the rock cohesion; P_p is the pore pressure; α is the effective stress coefficient of pore pressure; K is the coefficient related to the internal friction angle, defined as $K = \tan(45^\circ + \varphi/2)$; and η is a correction parameter, generally ranging from 0.95 to 1.

After the analytical solution of collapse pressure for each virtual well is obtained, it is uniformly converted into collapse pressure equivalent density to establish the dataset labels. A training dataset linking seismic records to collapse pressure is then constructed and can be used for learning and training deep neural networks. Figure 4 presents examples of the synthetic seismic envelope and collapse pressure equivalent density curves for three virtual wells. It can be observed that different virtual wells exhibit differences in local amplitude energy, interval-scale trends, and label-curve morphology, providing a sample basis for subsequent model learning of the mapping relationship between seismic attributes and wellbore stability parameters.

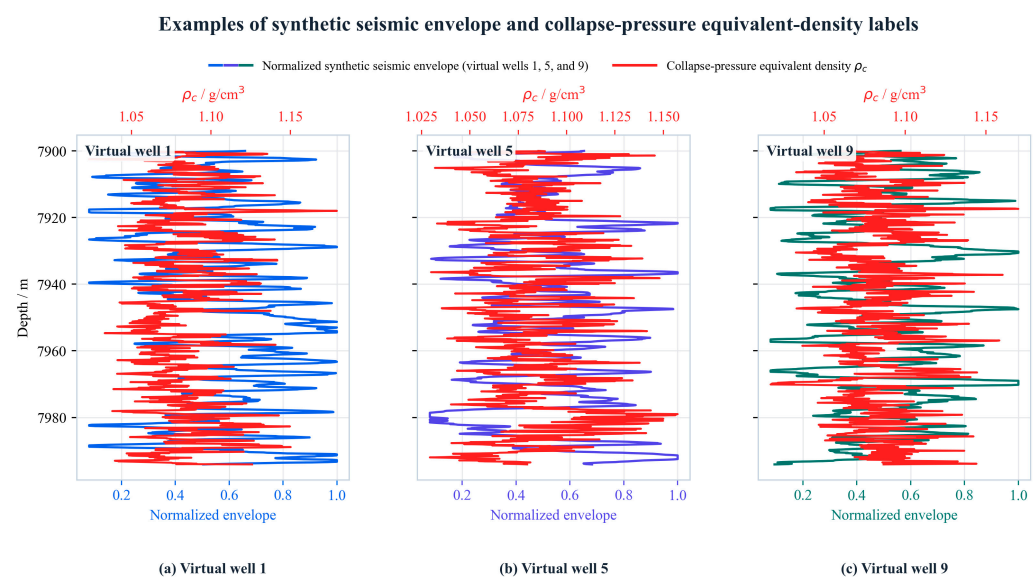


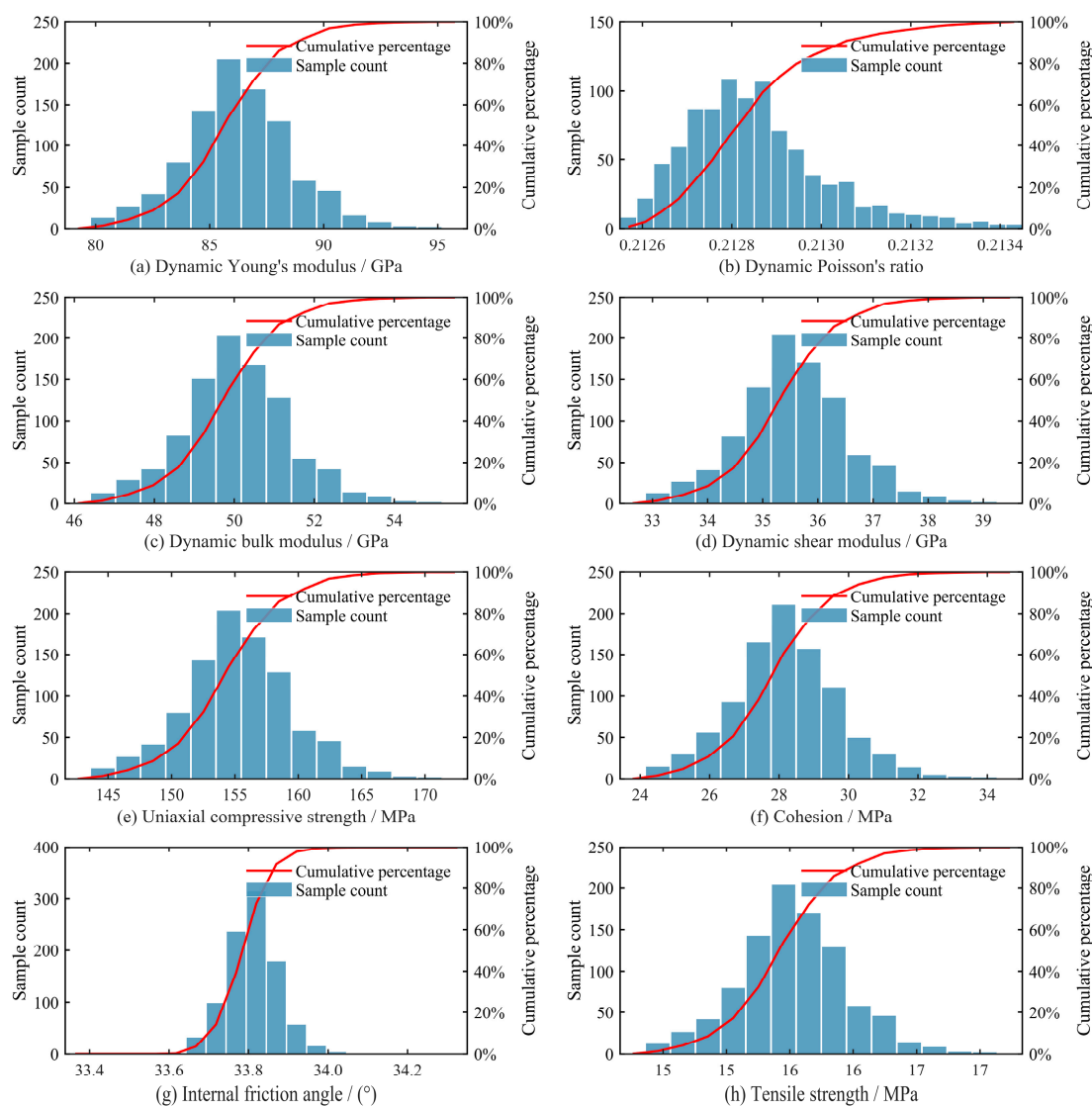
Figure 4. Examples of synthetic seismic attributes and collapse pressure equivalent density labels for virtual wells.

2.5. Dataset Construction

A total of five drilled wells are available in the study area. Four wells were used as conditioning wells for virtual-well generation, from which 1996 virtual-well profiles were constructed. Together with the four conditioning-well profiles, the model-development dataset therefore contained 2000 well sequences, each comprising 942 depth samples. The dataset was randomly divided into training and validation sets at a ratio of 7:3. Specifically, the training set contained 1400 wells, while the validation set contained 600 wells. In this study, each virtual well was treated as an independent sample unit, and data including the synthetic seismic envelope, acoustic impedance, and collapse pressure equivalent density were stored on a unified depth grid. The fifth well, which was not involved in training, was used as the prediction well to validate the model prediction results. An overview of the dataset is presented in Table 1 and Figures 5 and 6.

Table 1. Overview of the dataset.

Parameter	Minimum Value	Maximum Value
Dynamic Young's modulus/GPa	79.88	95.09
Dynamic Poisson's ratio	0.213	0.214
Dynamic bulk modulus/GPa	46.46	55.14
Dynamic shear modulus/GPa	32.91	39.21
Uniaxial compressive strength/MPa	143.80	171.17
Cohesion/MPa	24.22	34.21
Internal friction angle/°	33.64	34.05
Tensile strength/MPa	14.38	17.12

**Figure 5.** Distribution of rock mechanical parameters in actual wells.

The objective of data preprocessing is to remove outliers, unify data scales, and form sample inputs that can be trained stably. Data preprocessing mainly includes outlier processing, depth-grid unification, training-validation splitting, normalization or standardization, and model sample organization. First, invalid values were removed, and the synthetic seismic envelope and ρ_c were checked to ensure that they were within reasonable engineering ranges. Then, all virtual wells were unified onto the same depth grid, so that each well had input and output sequences with the same length.

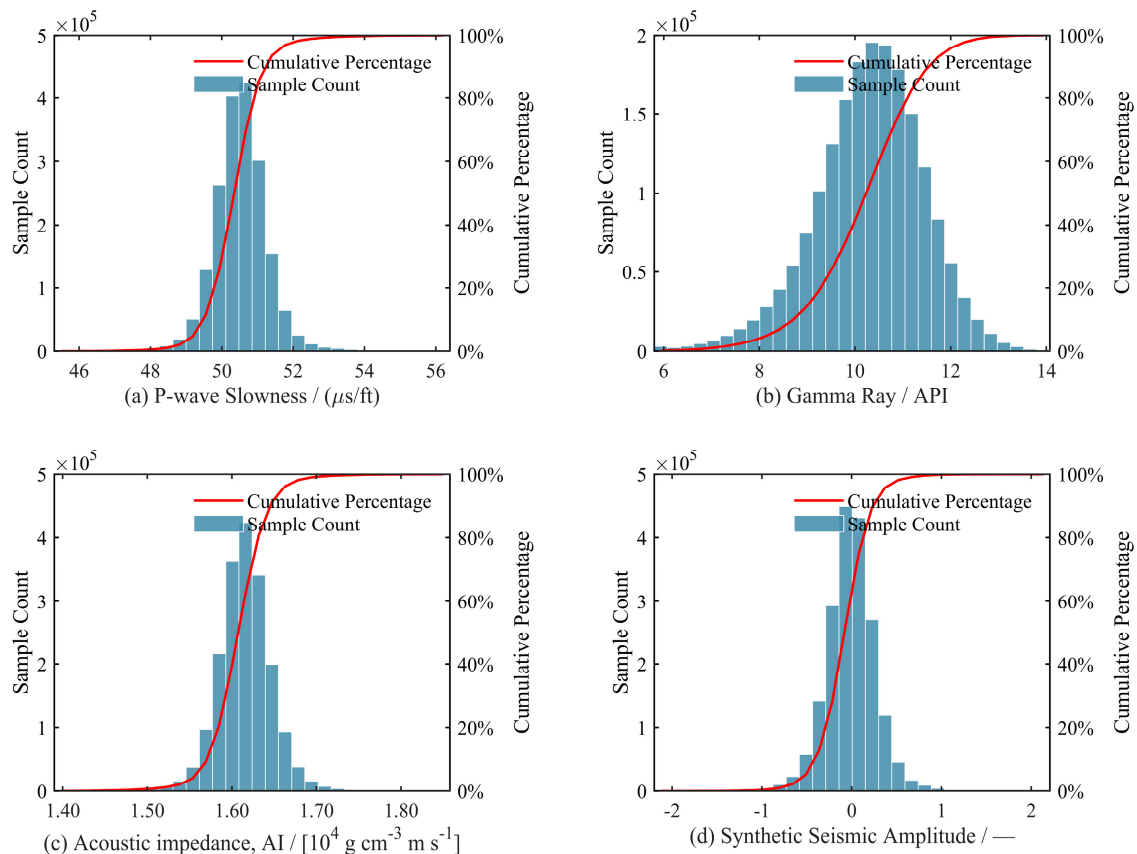


Figure 6. Distribution of parameters in virtual wells.

For traditional machine learning models, each depth point or local window is commonly treated as one training sample. Local-window features can retain information from adjacent depth points to a certain extent. For deep learning models, the sequence structure is preserved, and the input can be extended to multi-channel features, including the synthetic seismic envelope, envelope gradient, absolute gradient, and normalized depth. Let x_i be the original input variable. The commonly used min–max normalization can be expressed as:

$$x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (5)$$

When the model is more sensitive to the distribution of variables, standardization can be applied:

$$\tilde{x}_i = \frac{x_i - \mu}{\sigma + \varepsilon} \quad (6)$$

where μ and σ are the mean and standard deviation of the training set, respectively, and ε is a very small constant used to prevent the denominator from being zero. It should be noted that the normalization or standardization parameters must be estimated only from the training set and then applied to the validation set. These parameters should not be estimated using all samples together; otherwise, statistical information from the validation set would leak into the training process.

After the above processing, the dataset preserves the correspondence between the synthetic seismic attributes of virtual wells and the collapse pressure equivalent density labels, while also satisfying the requirements of machine learning models in terms of sample independence, numerical stability, and input format.

3. Machine Learning Models

3.1. Model Development Workflow

Machine learning models suitable for regression problems mainly include four categories: regression trees, linear regression, support vector machines, and neural networks. Considering the nonlinear mapping relationship between collapse pressure and the input parameters, four models suitable for nonlinear regression were selected from these categories for collapse pressure prediction, namely Random Forest, Polynomial Regression, Support Vector Regression, and CNN-MultiLSTM. These models were compared to identify the most suitable machine learning model.

The machine-learning and deep-learning models were implemented in Python 3.9.18 using PyTorch 1.9.1 and scikit-learn 1.5.1. GPU acceleration was provided by CUDA Toolkit 11.1 and cuDNN 8.0.5. Supporting packages included NumPy 1.26.4, pandas 1.4.4, SciPy 1.13.1, Matplotlib 2.2.5, and joblib 1.4.2.

3.2. Random Forest Regression Model

Random Forest is a nonparametric regression model based on an ensemble of decision trees. Its basic idea is to use the bootstrap method to draw multiple subsample sets with replacement from the training samples, train one regression tree on each subsample set, and randomly select a subset of features for optimal splitting at each node. This strategy reduces the sensitivity of a single tree to perturbations in training samples and input features. For collapse pressure equivalent density prediction, Random Forest can handle nonlinear relationships between input features and the prediction target, and is insensitive to feature scaling, making it an important baseline among traditional machine learning models.

Let the m th regression tree obtained after training be $T_m(x)$, and let the total number of trees be M . The prediction of the Random Forest for sample x can be expressed as the average output of all regression trees:

$$\hat{\rho}_c(x) = \frac{1}{M} \sum_{m=1}^M T_m(x) \quad (7)$$

where $\hat{\rho}_c(x)$ is the predicted collapse pressure equivalent density, and x is the input feature vector of the model. Random Forest has the advantage of strong robustness and can alleviate the overfitting problem commonly observed in individual decision trees. Meanwhile, because each tree uses only a subset of samples and a subset of features, the model has a certain tolerance to local noise. However, its limitation is that it is essentially a point-wise or local-window regression model and cannot directly represent long-range dependencies in the depth sequence of an entire well. Therefore, for predicting collapse pressure equivalent density curves with clear vertical trends, it is usually necessary to compare Random Forest with more powerful sequence models. The structure of the Random Forest model is illustrated in Figure 7.

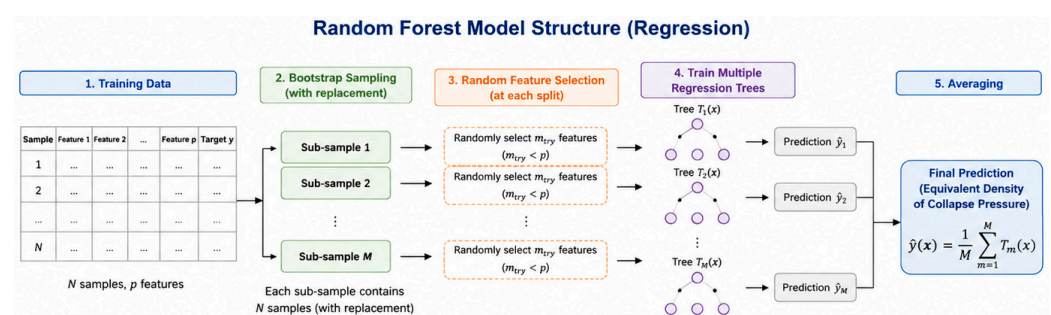


Figure 7. Schematic diagram of the Random Forest regression model.

3.3. Polynomial Regression Model

Polynomial Regression is an explicit nonlinear model extended from ordinary linear regression. Its basic idea is to map the original input variables into a high-dimensional feature space through polynomial functions, and then establish a linear regression relationship in that space. Let the input feature be x , and let the feature vector after polynomial expansion be denoted as $\phi(x)$. The Polynomial Regression model can be expressed as:

$$\hat{\rho}_c(x) = \beta_0 + \beta^T \phi(x) \quad (8)$$

When a second-order polynomial is used, $\phi(x)$ includes not only the original features, but also the squared terms and interaction terms of each feature. For example, for two input variables, x_1 and x_2 , the second-order polynomial features can be expressed as:

$$\phi(x) = [x_1, x_2, x_1^2, x_1x_2, x_2^2]^T \quad (9)$$

The advantage of Polynomial Regression is that its model form is clear and its parameters have relatively strong interpretability. Therefore, it can be used as a basic model to determine whether a simple nonlinear relationship exists between seismic attributes and collapse pressure equivalent density. To avoid overfitting caused by the increased dimensionality of polynomial features, Ridge regularization can usually be introduced. Its objective function is:

$$\min_{\beta} \sum_{i=1}^n (y_i - \beta_0 - \beta^T \phi(x_i))^2 + \lambda \|\beta\|_2^2 \quad (10)$$

where y_i is the true collapse pressure equivalent density of the i th sample, and λ is the regularization coefficient. For the dataset used in this study, Polynomial Regression can describe part of the curved trend between the amplitude envelope and the equivalent density. However, because its nonlinear order is limited and it lacks contextual information from the depth sequence, it is prone to underfitting in complex interval responses. Therefore, Polynomial Regression is more suitable as an explicit nonlinear baseline rather than as the final prediction model. The principle of Polynomial Regression is shown in Figure 8.

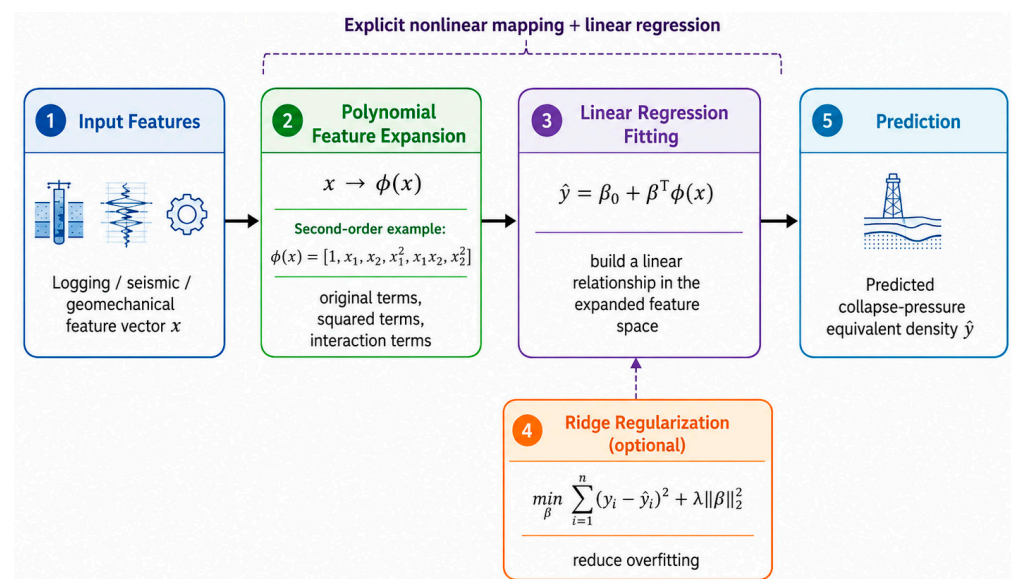


Figure 8. Schematic diagram of Polynomial Regression.

3.4. Support Vector Regression Model

Support Vector Regression (SVR) is a regression method developed from the classification concept of Support Vector Machine. Unlike ordinary least-squares regression, SVR introduces an insensitive loss function. When the prediction error is smaller than a given threshold ε , it is not included in the loss; only samples with errors exceeding ε are penalized in the objective function. This mechanism gives the model strong robustness to small-amplitude noise, making it suitable for regression problems with local perturbations but clear overall trends.

For linear SVR, the regression function can be expressed as:

$$f(x) = \mathbf{w}^T \mathbf{x} + b \quad (11)$$

The corresponding optimization objective can be written as:

$$\begin{aligned} y_i - f(x_i) &\leq \varepsilon + \xi_i, \quad i = 1, 2, \dots, n \\ f(x_i) - y_i &\leq \varepsilon + \xi_i^*, \quad i = 1, 2, \dots, n \\ \xi_i, \xi_i^* &\geq 0, \quad i = 1, 2, \dots, n \end{aligned} \quad (12)$$

where C is the penalty coefficient, and ξ_i and ξ_i^* are slack variables. When a nonlinear relationship exists between the input features and the prediction target, the input can be mapped into a high-dimensional space through a kernel function. Among them, the radial basis function kernel is commonly expressed as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2) \quad (13)$$

The advantage of SVR is that it has a well-established theoretical foundation and often exhibits good generalization capability for small- and medium-sized samples. However, for datasets consisting of millions of depth sampling points, full kernel SVR has a high computational cost, and kernel approximation methods are usually required to reduce the computational complexity. Figure 9 illustrates the basic principle of Support Vector Regression.

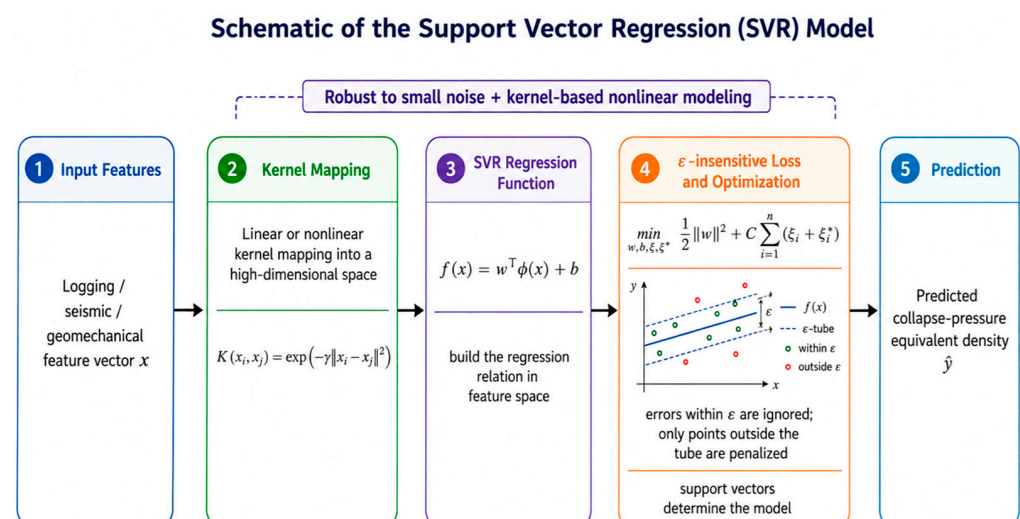


Figure 9. Schematic diagram of Support Vector Regression.

3.5. CNN-MultiLSTM Deep Learning Model

The CNN-MultiLSTM model is used to characterize local sequential variations in seismic attributes along the depth direction. This model generally takes the synthetic

seismic envelope window around the central depth point as input. First, local waveform features are extracted through a one-dimensional convolutional layer. Then, sequence dependencies within the window are captured by multiple LSTM layers. Finally, the collapse pressure equivalent density at the central depth point is output through a fully connected layer. Compared with traditional machine learning models, CNN-MultiLSTM does not rely entirely on manually designed statistical features, but can automatically extract patterns related to the prediction target from local sequences.

The computation of the one-dimensional convolutional layer can be expressed as follows [23]:

$$h_t^{(k)} = \sigma \left(\sum_{j=0}^{K-1} W_j^{(k)} x_{t+j} + b^{(k)} \right) \quad (14)$$

where K is the convolution kernel length, $W_j^{(k)}$ and $b^{(k)}$ are the weight and bias of the k th convolution kernel, respectively, and σ is the nonlinear activation function. The LSTM unit controls information transmission through the input gate, forget gate, and output gate [25,26]. Its core computational process can be written as:

$$\begin{aligned} f_t &= \sigma(W_f[h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i[h_{t-1}, x_t] + b_i) \\ \tilde{C}_t &= \tanh(W_C[h_{t-1}, x_t] + b_C) \\ C_t &= f_t C_{t-1} + i_t \tilde{C}_t \\ o_t &= \sigma(W_o[h_{t-1}, x_t] + b_o) \\ h_t &= o_t \tanh(C_t) \end{aligned} \quad (15)$$

where f_t , i_t , and o_t denote the forget gate, input gate, and output gate, respectively; C_t is the cell state; and h_t is the hidden state. After stacking multiple LSTM layers, the model can achieve enhanced representation capability for local depth sequences. For the collapse pressure equivalent density prediction problem addressed in this study, CNN-MultiLSTM can make use of waveform variations in the seismic envelope within a local depth range. Therefore, compared with Random Forest, Polynomial Regression, and Support Vector Regression, it has a stronger ability to extract sequential features. Figure 10 presents a schematic illustration of the basic principle of the CNN-MultiLSTM model.

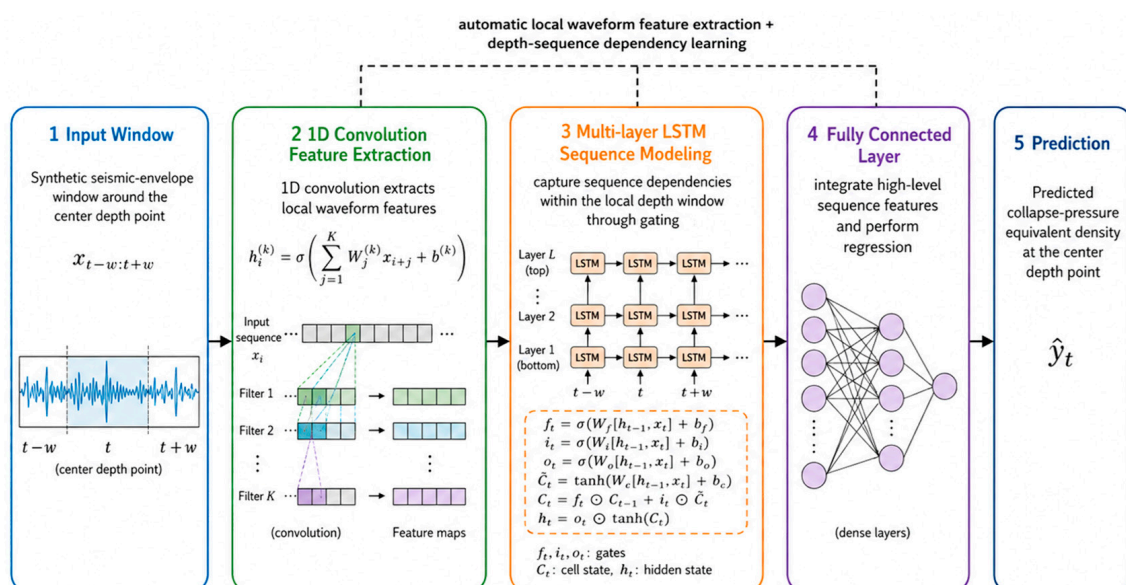


Figure 10. Schematic diagram of the CNN-MultiLSTM model.

3.6. Model Evaluation Metrics

To quantitatively evaluate the prediction performance of different models, this study adopts the mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination (R^2) as the main evaluation metrics. Suppose that there are n samples in the validation set, y_i is the true collapse pressure equivalent density of the i th sample, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of the true values. The metrics are defined as follows:

$$\begin{aligned} MAE &= \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \\ RMSE &= \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \\ MAPE &= \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \\ R^2 &= 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \end{aligned} \quad (16)$$

MAE directly reflects the average absolute deviation between the predicted values and the true values. Its unit is the same as that of the collapse pressure equivalent density, making it convenient for interpreting model errors from an engineering perspective. RMSE is more sensitive to larger errors and can indicate whether the model exhibits obvious prediction distortion in local depth intervals. MAPE expresses the relative error in percentage form, which facilitates comparison of error levels across different data ranges. R^2 reflects the ability of the model to explain the variance of the target variable. The closer its value is to 1, the better the overall fitting performance of the model.

4. Selection and Optimization of Machine Learning Models

4.1. Model Selection Scheme

After the construction of virtual well data, extraction of synthetic seismic attributes, and standardization of training samples, it is necessary to further determine the optimal machine learning model for collapse pressure equivalent density prediction. Since the relationship between collapse pressure equivalent density and the seismic amplitude envelope is not simply linear, and the depth direction is jointly affected by continuous sedimentary intervals, local waveform variations, and the overall compaction trend, model selection should not rely on a single algorithm. Instead, different types of models should be compared under a unified dataset and unified evaluation metrics. To ensure the comparability of the results, this study uses the virtual well samples established above. Five actual drilled wells were available. Four wells were used as conditioning wells for virtual-well generation, from which 1996 virtual-well profiles were constructed. Together with the four conditioning-well profiles, the model-development dataset therefore contained 2000 well sequences, each comprising 942 depth samples. The 2000 sequences were divided at the well level into 1400 training wells and 600 validation wells. Each well contains equally spaced sampling points within the depth range of 7900.000–7994.100 m, and the validation set contains a total of 565,200 depth samples.

The candidate models include four types: Random Forest, Polynomial Regression, Support Vector Regression, and CNN-MultiLSTM. These models cover different modeling concepts, including ensemble learning, explicit function fitting, kernel methods, and deep sequence learning, which helps evaluate the applicability of deep sequence models to this

problem from a methodological perspective. Model evaluation was conducted using four metrics: mean absolute error, root mean square error, mean absolute percentage error, and coefficient of determination. To avoid the influence of differences in random sampling on model ranking, the same training and validation set split was used for all four types of models, and the validation set metrics were used as the basis for model selection.

4.2. Comparison of Prediction Performance Among Different Machine Learning Models

As shown in Figure 11, the four types of machine learning models exhibit clear differences in prediction accuracy on the training, validation, and test sets. The prediction errors of the Random Forest model are generally at an intermediate level, with an MAE of approximately 0.0080–0.0087 g/cm³, an RMSE of approximately 0.0103–0.0113 g/cm³, a MAPE of approximately 0.70–0.80%, and an R² of approximately 0.77–0.79. This indicates that Random Forest can characterize the nonlinear relationship between seismic responses and collapse pressure to a certain extent, but its ability to represent continuous variations along the depth direction remains limited. Polynomial Regression and Support Vector Regression show relatively larger prediction errors. Specifically, Polynomial Regression has an MAE of approximately 0.0097 g/cm³ and an RMSE of approximately 0.0124 g/cm³, while Support Vector Regression has an MAE of approximately 0.0100 g/cm³, an RMSE of approximately 0.0128 g/cm³, a MAPE of approximately 0.90%, and an R² of approximately 0.69–0.72. These results indicate that relying solely on explicit function fitting or kernel-function mapping is insufficient to fully describe the complex nonlinear relationship among synthetic seismic data, depth-sequence features, and collapse pressure.

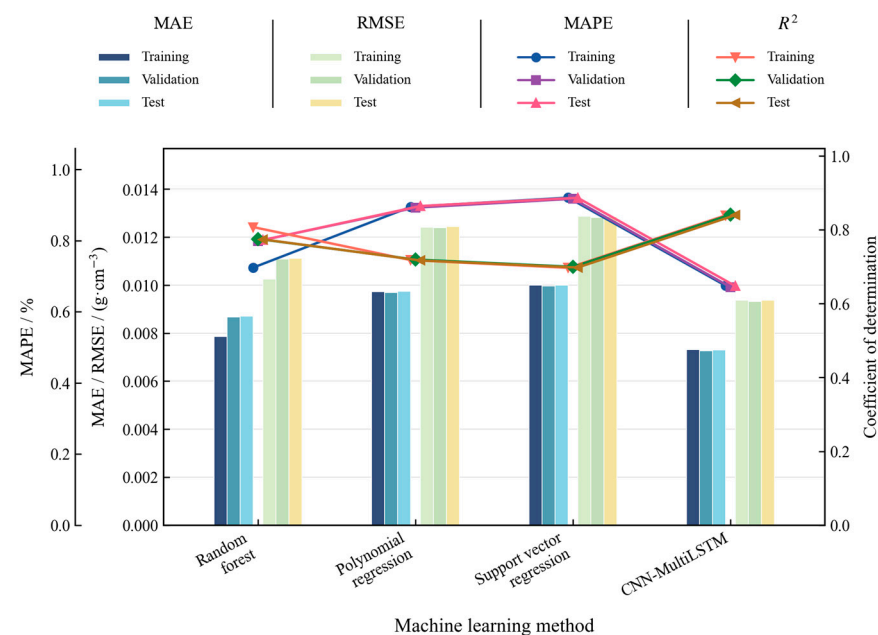


Figure 11. Comparison of prediction accuracy of different machine learning models on the training, validation, and test sets.

In contrast, the CNN-MultiLSTM model achieves higher prediction accuracy and better stability on the training, validation, and test sets. Its MAE is approximately 0.0073 g/cm³, RMSE is approximately 0.0093–0.0095 g/cm³, MAPE is approximately 0.64–0.67%, and R² is approximately 0.88–0.89. Compared with the Random Forest model, CNN-MultiLSTM further reduces both MAE and RMSE and significantly improves R². Compared with Polynomial Regression and Support Vector Regression, the improvement in prediction accuracy is even more pronounced. This indicates that the CNN-MultiLSTM model can simultaneously utilize the ability of convolutional layers to extract local seismic waveform

features and the ability of LSTM layers to learn sequence dependencies along the depth direction, thereby more effectively capturing the intrinsic relationship between synthetic seismic responses and collapse pressure. Therefore, the CNN-MultiLSTM model exhibits higher accuracy, stability, and generalization capability in collapse pressure prediction, and can be selected as the preferred model for subsequent prediction analysis.

As shown in Figure 12, clear differences can be observed in the scatter relationships between the true and predicted values on the validation set for the four types of models. Although the scatter points of the traditional machine learning models are generally distributed along the diagonal line, a certain shrinkage phenomenon occurs at both the high and low equivalent-density ends, indicating that these models tend to predict extreme samples toward the mean. In contrast, the scatter band of CNN-MultiLSTM is closer to the ideal diagonal line, especially showing better continuity and smaller dispersion in the medium- to high-equivalent-density range. This indicates that the deep sequence model not only improves the overall error metrics, but also enhances the ability to characterize the variation amplitude of collapse pressure equivalent density across different depth intervals. Therefore, when the virtual-well samples exhibit clear depth-sequence characteristics, CNN-MultiLSTM is more suitable as a collapse pressure equivalent density prediction model than point-wise or weakly sequential models.

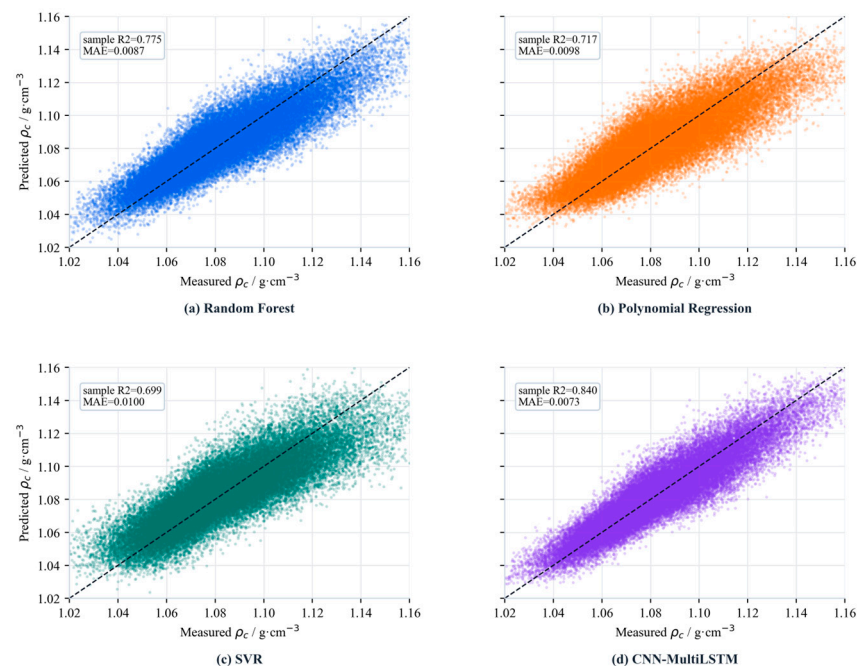


Figure 12. Relationship between true and predicted values on the validation set for different models.

4.3. Structural and Hyperparameter Optimization of the Whole-Well Sequence Model

After CNN-MultiLSTM was identified as the preferred baseline model, parameter optimization was further conducted with respect to the number of training samples, local window length, and input feature channels. The baseline CNN-MultiLSTM uses the synthetic seismic envelope as a single input channel. A local window consisting of 17 sampling points around the central depth point is used as the model input. Local waveform features are first extracted through a one-dimensional convolutional layer, and sequence dependencies within the window are then established by multiple LSTM layers. Finally, the collapse pressure equivalent density of the central point is output.

Training sample size optimization was first performed. When the number of training samples was increased from 300,000 to 600,000 and the batch size was adjusted to 8192, the model R² showed a slight improvement. This result indicates that, under the original

local-window structure and single-envelope input condition, simply increasing the number of training samples can slightly improve model stability, but cannot fundamentally break through the upper limit of the model representation capability. This is because strong correlations exist among the virtual well samples. When the dimensionality of the model input information remains unchanged, additional samples mainly improve the stability of parameter estimation, but are unlikely to provide new depth-context information.

Local window length optimization was then performed. When the window length was increased from 17 sampling points to 25 sampling points, the model R^2 further improved. Window expansion enables the model to observe waveform variations over a wider depth range, which helps identify the slowly varying trends and short-period perturbations of the amplitude envelope within local intervals. However, even with an increased window length, the model still performs local point-wise prediction. The available depth range remains limited, making it unable to fully characterize the background trend from top to bottom across the entire virtual well. Therefore, window optimization can bring limited improvement, but the magnitude of improvement is not significant.

Input channel optimization was finally performed. After normalized depth was added as the second input channel on the basis of the synthetic seismic envelope, the model R^2 increased to 0.8579, and the improvement was significantly greater than that achieved by simply increasing the number of samples or expanding the window length. This result indicates that the collapse pressure equivalent density is controlled not only by local seismic amplitude, but also by depth position and the compaction background. The normalized depth channel provides the model with an absolute position constraint, enabling it to distinguish different collapse pressure levels corresponding to similar amplitude patterns at different depths, thereby significantly improving prediction accuracy.

The above three groups of experiments show that increasing the sample size, expanding the local window, and introducing the depth channel can all improve the CNN-MultiLSTM model to a certain extent. However, the magnitude of improvement remains limited, indicating that further performance gains depend not only on parameter adjustment, but also on enhancing the model's ability to represent long-range depth dependencies. Therefore, the optimization strategy was further extended from local-window point-wise prediction to whole-well sequence modeling.

On the basis of the sequence-learning framework established above, the model structure was further optimized by introducing residual dilated one-dimensional convolutions. The dilated convolution enlarges the receptive field along the depth direction, while the residual connections facilitate stable training of the deeper sequence network. Combined with whole-well sequence input and whole-well curve output, this structure enables the model to utilize both local waveform variations and long-range depth-context information within a unified prediction framework.

The optimized model adopts a residual dilated one-dimensional convolutional sequence network. The input features include four channels: the synthetic seismic envelope, the first-order gradient of the envelope, the absolute value of the gradient, and normalized depth. Among them, the envelope channel reflects seismic amplitude energy, the gradient channel represents the direction of amplitude variation with depth, the absolute gradient characterizes the intensity of local variations, and normalized depth provides a positional constraint. During training, each virtual-well sequence contains 942 depth sampling points corresponding to the unified target interval of 7900.000–7994.100 m. It should be noted that 942 represents the sequence length of the training samples rather than a fixed input dimension imposed by the network architecture. The final residual dilated convolutional model is fully convolutional along the depth dimension. Therefore, it can accept continuous sequences of different lengths during inference and generate output sequences with the

corresponding length. The model takes the complete multi-channel sequence as input and simultaneously predicts the collapse-pressure equivalent-density profile along the input depth grid. Compared with the local point-wise model, the whole-well sequence model can utilize contextual depth information within the same forward propagation process, allowing the prediction results to maintain better consistency between the overall trend and local details.

The core advantage of the residual dilated convolution structure is that it can enlarge the receptive field with relatively few parameters. Let the input sequence be x_t , the convolution kernel length be K , and the dilation rate be d . The dilated convolution can then be expressed as:

$$y_t = \sum_{j=0}^{K-1} w_j x_{t-d \cdot j} \quad (17)$$

As the dilation rate increases layer by layer, the model can simultaneously utilize information from both adjacent depth points and distant depth points. The residual connection can be expressed as follows:

$$z_{l+1} = z_l + F(z_l, \theta_l) \quad (18)$$

where $F(\cdot)$ denotes the nonlinear transformation in the l th residual block. The residual connection can alleviate the gradient decay problem during deep network training, enabling the model to maintain stable optimization while deepening the network structure and enlarging the receptive field.

As shown in Figure 13, the prediction accuracy of the model optimized using whole-well sequences is greatly improved. This indicates that, compared with the local point-wise model, the whole-well sequence model can utilize contextual depth information within the same forward propagation process, allowing the prediction results to maintain better consistency between the overall trend and local details.

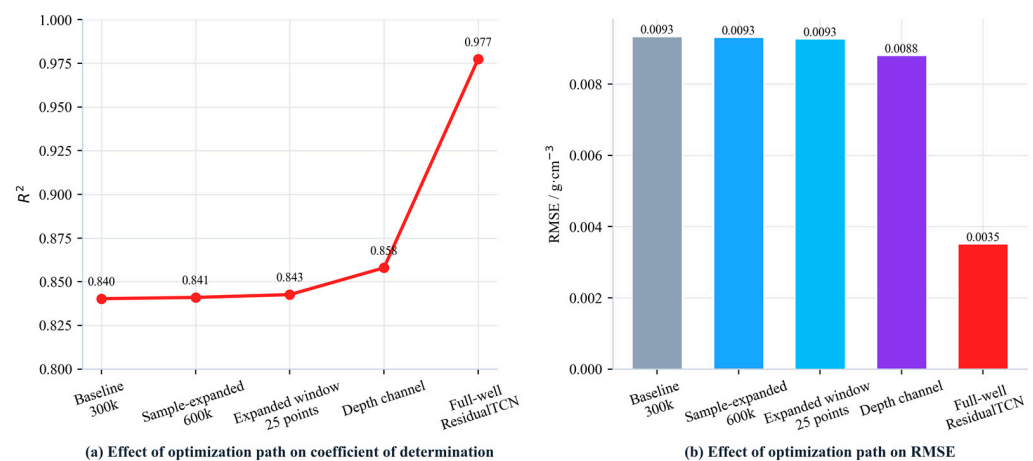


Figure 13. Structural optimization path from the CNN-MultiLSTM model to the whole-well sequence model.

After the whole-well sequence structure was determined as the optimal structure, hyperparameter optimization was carried out based on the whole-well sequence model. Referring to the balanced combination concept of orthogonal experiments, this study selected the number of hidden channels, number of residual blocks, convolution kernel length, Dropout ratio, and learning rate as the main optimization factors of the whole-well sequence model. Four levels were set for each factor, forming an $L16(4^5)$ hyperparameter optimization design scheme, as shown in Table 2. During model selection, the validation-

set R^2 was used as the governing criterion to evaluate the overall fitting capability of the predicted collapse-pressure profile. RMSE was used as a secondary error metric, while MAE and MAPE were also considered to assess the absolute and relative error levels.

Table 2. Statistical results of hyperparameter optimization for the whole-well sequence model.

No.	Number of Hidden Channels	Number of Residual Blocks	Convolution Kernel Length	Dropout Ratio	Learning Rate	Best Epoch	RMSE (g/cm ³)	R ²
1	64	8	3	0.00	5×10^{-4}	149	0.0043	0.9658
2	64	12	5	0.05	1×10^{-3}	134	0.0032	0.9733
3	64	16	7	0.10	2×10^{-3}	160	0.0037	0.9767
4	64	20	9	0.20	5×10^{-3}	115	0.0031	0.9734
5	96	8	5	0.10	5×10^{-3}	132	0.0032	0.9760
6	96	12	3	0.20	2×10^{-3}	172	0.0031	0.9762
7	96	16	9	0.00	1×10^{-3}	212	0.0037	0.9740
8	96	20	7	0.05	5×10^{-4}	215	0.0035	0.9742
9	128	8	7	0.20	1×10^{-3}	136	0.0035	0.9774
10	128	12	9	0.10	5×10^{-4}	104	0.0036	0.9741
11	128	16	3	0.05	5×10^{-3}	49	0.0048	0.9664
12	128	20	5	0.00	2×10^{-3}	134	0.0038	0.9739
13	160	8	9	0.05	2×10^{-3}	54	0.0035	0.9743
14	160	12	7	0.00	5×10^{-3}	47	0.0038	0.9711
15	160	16	5	0.20	5×10^{-4}	165	0.0034	0.9771
16	160	20	3	0.10	1×10^{-3}	162	0.0036	0.9768

According to the results of the 16 combinations shown in Table 2, combinations 4 and 6 achieved the minimum rounded RMSE of 0.0031 g/cm³, with R^2 values of 0.9734 and 0.9762, respectively. Combination 9 was selected because it achieved the highest R^2 of 0.9774 among all combinations while maintaining a relatively low RMSE of 0.0035 g/cm³. Therefore, R^2 governed the final selection, whereas RMSE and the other error metrics were used as supporting indicators. The selected combination consists of 128 hidden channels, 8 residual dilated convolution blocks, a convolution kernel length of 7, a Dropout ratio of 0.20, and a learning rate of 0.001. A weight decay coefficient of 1×10^{-5} , a batch size of 64, and a maximum training epoch of 220 were used. Early stopping was applied based on the validation-set error, and the optimal model appeared at epoch 136. This model achieved an MAE of 0.00262 g/cm³, an RMSE of 0.00352 g/cm³, a MAPE of 0.241%, and an R^2 of 0.9774 on the validation set, showing a significant improvement over the baseline CNN-MultiLSTM model.

From the optimization process, increasing the sample size, expanding the window length, and extending the input channels can all improve model performance to a certain extent. However, the most significant improvement comes from the whole-well sequence structure. The advantage of the baseline CNN-MultiLSTM lies in its ability to process local sequences, but its prediction target remains a single-point label. In contrast, the optimized whole-well sequence model extends the prediction object to the complete curve, enabling the model to simultaneously focus on the overall consistency among multiple depth points at the loss-function level. Therefore, the optimization logic in this section is not to tune hyperparameters independently before replacing the model structure, but rather to first determine the whole-well sequence as the optimal structure through stepwise experiments, and then identify the final hyperparameter combination based on the whole-well sequence model. The final model further enhances the ability to represent long-range depth dependencies and impose overall curve constraints on the basis that CNN-MultiLSTM has already demonstrated the effectiveness of deep sequence learning.

4.4. Internal Validation and Analysis of the Optimal Model

As shown in Figure 14, the training process of the optimal model, the relationship between true and predicted values on the validation set, and the residual distribution all exhibit good stability. The training loss and validation loss decrease rapidly in the early stage and then gradually converge, indicating that the model can effectively learn

the mapping relationship between seismic attributes and collapse pressure equivalent density, without showing obvious overfitting characterized by decreasing training error and increasing validation error. The scatter points on the validation set are concentrated near the ideal diagonal line, and the predicted values cover the low-, medium-, and high-value ranges, indicating that the model has good fitting capability for different collapse pressure levels. The residual histogram shows a narrow distribution centered around zero, suggesting that the model errors are mainly concentrated within a small range and that the prediction results do not exhibit obvious systematic overestimation or underestimation.

Training and Validation Results of the Optimal ResidualTCN Full-Well Sequence Model

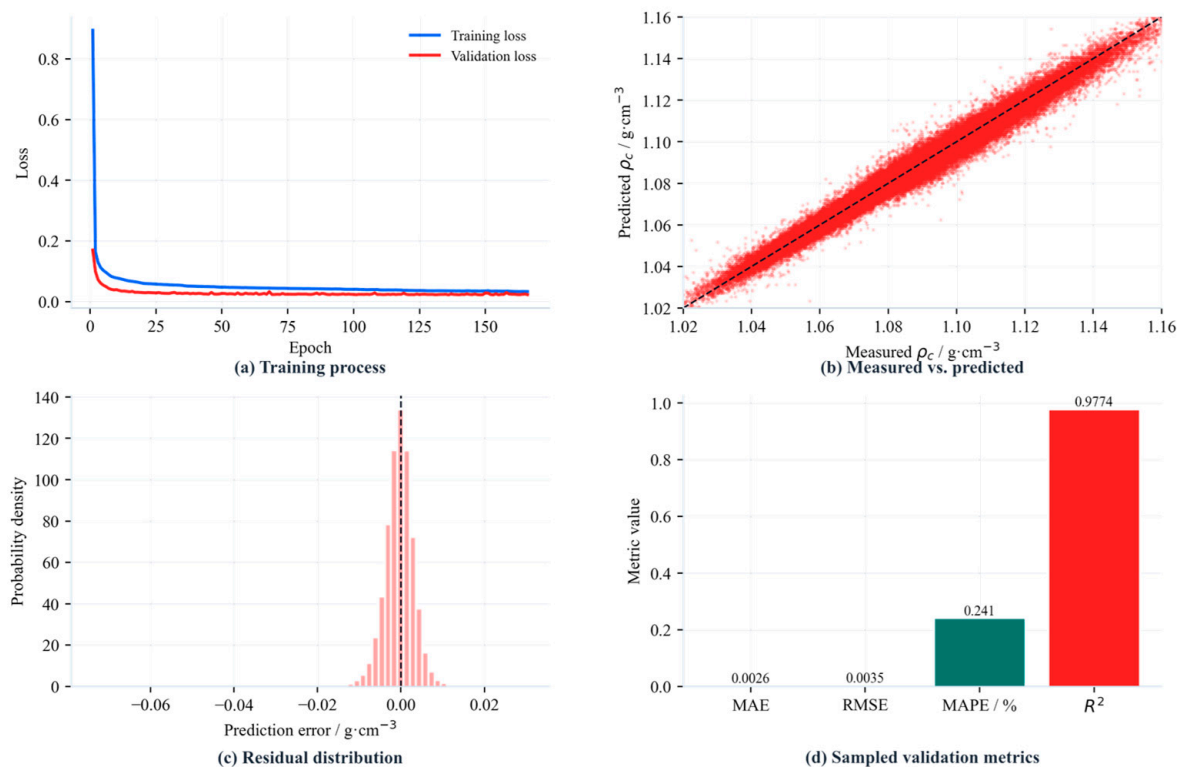


Figure 14. Training process of the optimal model and prediction performance on the validation set.

As shown in Figure 15, the true collapse pressure equivalent density curve of the typical validation well is in good agreement with the curve predicted by the optimal model. The optimal model can not only fit the overall trend with depth, but also effectively preserve the fluctuation patterns in local intervals. The local curve oscillation and extreme-value regression problems that commonly occur in traditional point-wise models are significantly improved by the whole-well sequence model through the shared use of contextual information from the entire well. This result indicates that the optimal model has acquired the ability to predict continuous collapse pressure equivalent density profiles from synthetic seismic responses, providing a data basis for subsequent collapse pressure prediction at target well locations and pre-drilling analysis of the safe mud density window.

Based on the results of model selection, structural selection, and hyperparameter optimization for the whole-well sequence model, traditional machine learning models can establish a preliminary mapping between seismic attributes and collapse pressure equivalent density. However, their validation accuracy is relatively limited due to constraints in input feature representation and depth-sequence modeling capability. CNN-MultiLSTM improves the extraction of local sequential features through the combined structure of convolutional layers and LSTM layers, and achieves the best performance among the four

baseline models, indicating that the proposed modeling framework is reasonable. On this basis, the stepwise increase in sample size, window expansion, and introduction of the depth channel further demonstrate that the improvement potential of the local point-wise model is limited. After the model was changed to whole-well sequence input and whole-well curve output, hyperparameters including the number of hidden channels, number of residual blocks, convolution kernel length, Dropout ratio, and learning rate were further optimized. Finally, the validation-set R^2 was increased to 0.9774, while the MAE and RMSE were reduced to 0.00262 g/cm³ and 0.00352 g/cm³, respectively. Therefore, the residual dilated convolutional whole-well sequence model is determined to be the optimal model for predicting virtual-well collapse pressure equivalent density in this study, and is also adopted as the final model for subsequent prediction analysis.

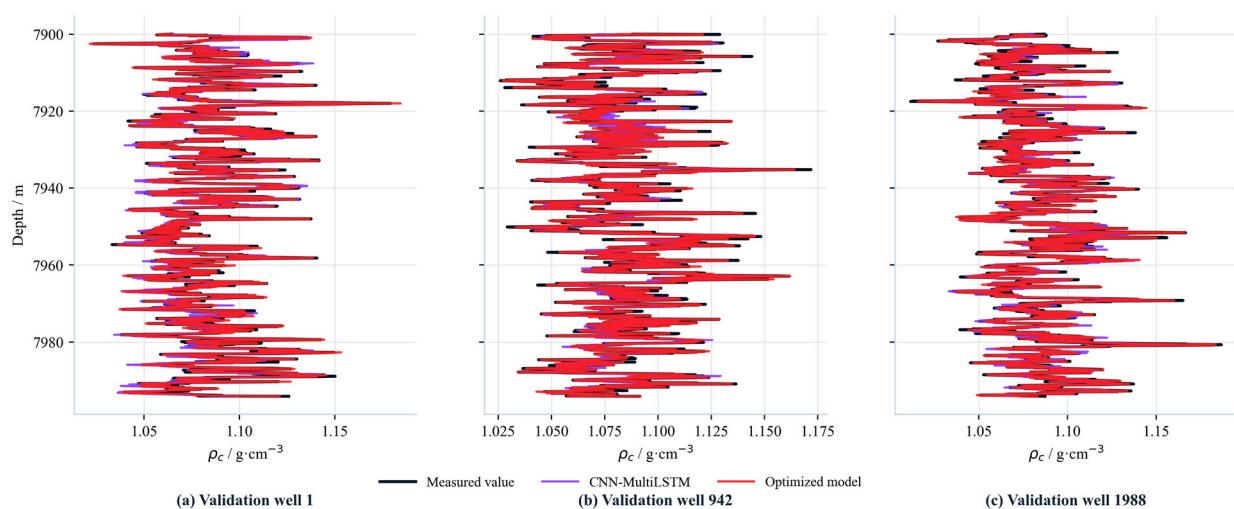


Figure 15. Predicted collapse pressure equivalent density curve for a typical validation well.

It should be emphasized that both the synthetic seismic inputs and the collapse-pressure labels are derived from the same underlying virtual-well acoustic-impedance realizations. The synthetic seismic responses are generated from impedance-derived reflection coefficients through wavelet convolution, whereas the collapse-pressure labels are calculated from impedance-related rock-mechanical parameters. Therefore, the high accuracy obtained on the internal validation set primarily reflects the model's ability to learn the nonlinear mapping embedded in this physically constrained synthetic-data framework and should not be interpreted directly as field-scale predictive accuracy. Independent validation using actual wells is therefore required to provide a more conservative and practically relevant assessment of model generalization. Accordingly, the following section evaluates the optimized model using actual well data that were not involved in model training.

5. Model Validation

5.1. Validation Strategy and Data Sources

To evaluate the generalization capability of the optimized model beyond the synthetic virtual-well domain, actual well data that were not involved in model training were used for external validation, this study selects actual Well X, which was not involved in training or model selection, as the external validation object. For the extended actual-well interval of 7770.888–7994.100 m, the 2238 sampling points were organized as a continuous input sequence and directly processed by the fully convolutional model using the same feature-construction and normalization procedures as those used for the training data. The consistency between the predicted curve and the actual curve is compared, and the MAE, RMSE, MAPE, and R^2 of different intervals are further calculated. In addition, the

variation of error with depth, statistical error distribution, and identification results of high-collapse-pressure intervals are analyzed. This enables a comprehensive evaluation of the model’s overall fitting capability, response capability to local abrupt changes, and engineering risk identification capability.

In this study, the prediction error is defined as the difference between the model-predicted collapse pressure equivalent density and the actual collapse pressure equivalent density, while the relative error is defined as the ratio of the absolute error to the actual value. MAE is used to characterize the average absolute deviation, RMSE is more sensitive to large local errors, MAPE is used to describe the relative error level, and R^2 is used to reflect the ability of the predicted values to explain the variance of the actual variations.

5.2. Profile Comparison of Collapse Pressure Prediction Results

Figure 16 presents the multi-parameter profile of the actual well, together with the comparison between the model-predicted collapse pressure and the actual collapse pressure. The last track in the figure shows the comparison of collapse pressure equivalent density curves, where the black curve represents the actual values and the red curve represents the model-predicted values. It can be seen that the predicted curve is in good overall agreement with the actual curve, and the model is able to reflect the main trend of collapse pressure variation with depth.

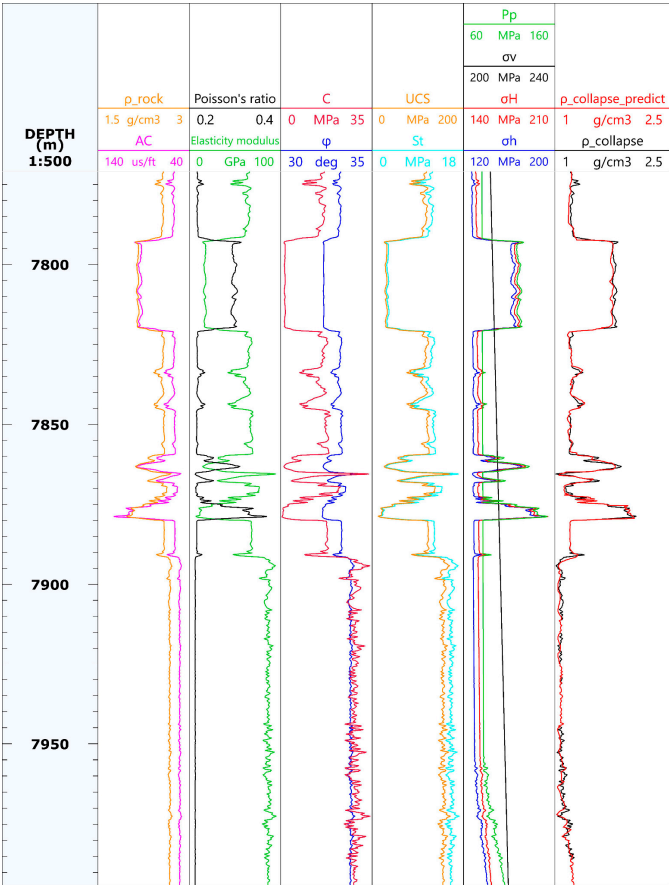


Figure 16. Comparison between model-predicted collapse pressure and actual collapse pressure.

In terms of profile morphology, obvious high-value intervals of actual collapse pressure occur near 7792.786–7819.800 m and 7875.200–7879.400 m. The predicted curve can simultaneously produce high responses in these intervals, indicating that the model has good identification capability for high-collapse-pressure intervals. Within the target interval below 7900 m, the overall fluctuation amplitude of the actual curve decreases, and the

predicted results also remain relatively stable, without persistent systematic overestimation or underestimation.

It should be noted that certain local deviations still exist between the predicted curve and the actual curve in intervals where the collapse pressure increases or decreases rapidly. These deviations are mainly concentrated in high-value perturbation intervals and may be related to the limited frequency bandwidth of seismic data, abrupt changes in local lithology or mechanical parameters, time–depth conversion errors, and the smoothing effect of the model on abrupt change points. Therefore, qualitative judgment based only on profile comparison is still insufficient, and further analysis using quantitative metrics and error distribution is required.

5.3. Prediction Accuracy Evaluation for Different Intervals

To quantify the prediction accuracy of the model, MAE, RMSE, MAPE, and R^2 were calculated separately for the full-well comparison interval, the target validation interval, the high-value perturbation interval, and the stable terminal interval. The results are shown in Table 3. Interval-based statistics can prevent a single global metric from masking error differences among local intervals and are also more consistent with the focus on key risk intervals in wellbore stability evaluation.

Table 3. Prediction accuracy evaluation of collapse pressure for different intervals.

Statistical Interval	Depth Interval/m	Number of Samples	MAE/(g/cm ³)	RMSE/(g/cm ³)	MAPE/%	R^2
Target validation interval	7900.000~7994.100	942	0.0182	0.0278	1.640	0.7356
Full-well comparison interval	7770.888~7994.100	2238	0.0246	0.0410	1.906	0.9712
High-value perturbation interval	7860.000~7880.000	201	0.0599	0.0911	4.073	0.8999
Stable terminal interval	7980.000~7994.100	142	0.0081	0.0165	0.681	0.6523

As shown in Table 3, the primary external validation was conducted within the target interval of 7900.000–7994.100 m. In this interval, the model achieved an MAE of 0.0182 g/cm³, an RMSE of 0.0278 g/cm³, a MAPE of 1.640%, and an R^2 of 0.7356. These results indicate that the model can reproduce the main depth-dependent variation in collapse pressure within the target interval, although local deviations remain. For the full-well comparison interval of 7770.888–7994.100 m, the model achieved an MAE of 0.0246 g/cm³, an RMSE of 0.0410 g/cm³, a MAPE of 1.906%, and an R^2 of 0.9712. The higher R^2 over the full-well comparison interval reflects the larger variation range of the actual collapse-pressure curve, whereas the larger MAE and RMSE are mainly associated with local high-value perturbations above the target interval.

In the high-value perturbation interval of 7860.000–7880.000 m, the MAE and RMSE were 0.0599 g/cm³ and 0.0911 g/cm³, respectively, indicating relatively large local errors where the collapse pressure changes rapidly. Nevertheless, the R^2 of this interval reached 0.8999, suggesting that the predicted curve still captured the main increasing and decreasing trends of the actual curve. In the stable terminal interval of 7980.000–7994.100 m, the MAE was 0.0081 g/cm³ and the MAPE was 0.681%, indicating relatively small prediction errors under a stable pressure background.

The lower R^2 values in the target validation interval and the stable terminal interval, compared with the full-well comparison interval, do not necessarily indicate larger prediction errors. In these intervals, the actual collapse pressure varies within a relatively narrow range, resulting in a low variance of the true values, to which R^2 is sensitive. Therefore, MAE, RMSE, MAPE, and profile morphology should be considered comprehensively rather than relying solely on R^2 when evaluating model performance in relatively stable intervals.

5.4. Variation of Error with Depth and Distribution Characteristics

To further analyze the distribution characteristics of prediction deviations along the depth direction, the curve of prediction error varying with depth was plotted, as shown in Figure 17. A positive error indicates that the model-predicted value is higher than the actual value, whereas a negative error indicates that the model-predicted value is lower than the actual value.

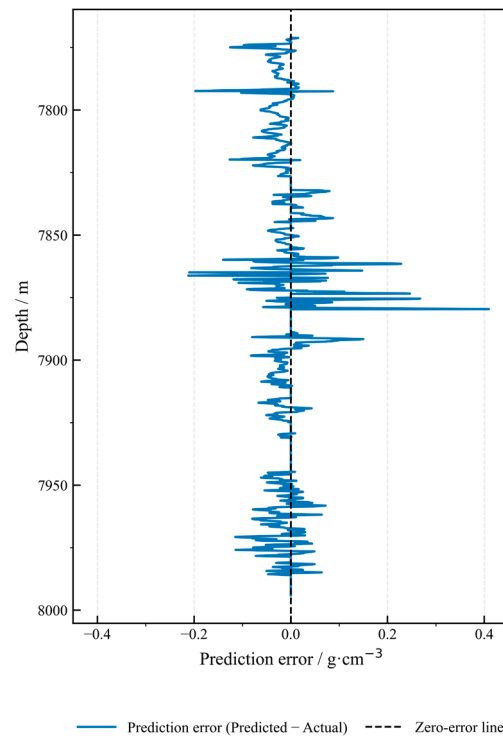


Figure 17. Variation of prediction error with depth.

As shown in Figure 17, the errors at most depth points are distributed around the zero-error line, and most errors fall within the range of $\pm 0.05 \text{ g/cm}^3$. Larger errors mainly occur in local perturbation intervals where the collapse pressure increases or decreases rapidly, especially near 7860–7880 m. This indicates that the model has good tracking capability for continuous and stable intervals, but is affected by the resolution of the input seismic responses and the smoothing characteristics of the model in local abrupt-change intervals.

Figure 18 shows the cumulative distribution characteristics of the absolute prediction error of the model. It can be observed that the cumulative distribution curve rises rapidly in the small-error range, indicating that the prediction errors of most samples are concentrated within a relatively low range. The mean absolute error of the model is 0.0246 g/cm^3 , and the median absolute error is 0.0162 g/cm^3 . The median is smaller than the mean, indicating that the model achieves high prediction accuracy for most samples, while a small number of samples with larger errors have a certain influence on the mean error. Overall, the absolute errors of the vast majority of samples are controlled within a small range, suggesting that the established model has good prediction stability and engineering applicability.

Table 4 further lists the proportions of errors falling within specified ranges for different intervals. As shown in Table 4, in the full-well comparison interval, the proportion of samples with $|e| \leq 0.05 \text{ g/cm}^3$ is 87.18%, and the proportion of samples with a relative error $\leq 5\%$ is 91.87%, indicating that the prediction results for most depth points fall within a small error range. In the target validation interval, the proportion of samples with $|e| \leq 0.05 \text{ g/cm}^3$ further increases to 92.36%, and the proportion with a relative error $\leq 5\%$

reaches 93.84%, indicating that the model has stable prediction capability within the key interval investigated in this study.

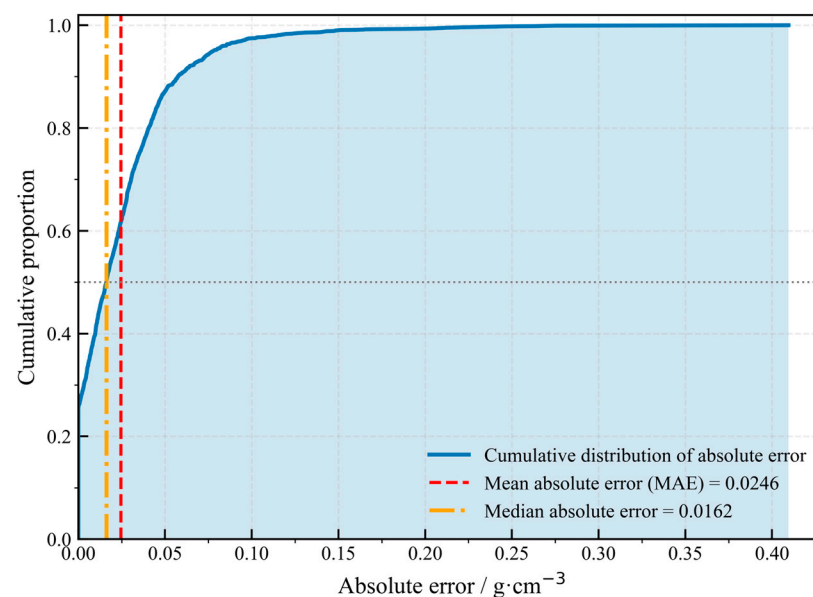


Figure 18. Statistical distribution of absolute prediction error.

Table 4. Error distribution statistics for different intervals.

Statistical Interval	$ e \leq 0.02\%$	$ e \leq 0.05\%$	Relative Error $\leq 2\%/%$	Relative Error $\leq 5\%/%$	95th-Percentile Absolute Error/(g/cm ³)	Maximum Absolute Error/(g/cm ³)
Full-well comparison interval	55.41	87.18	62.60	91.87	0.0778	0.4095
Target validation interval	61.04	92.36	64.01	93.84	0.0595	0.1150
High-value perturbation interval	27.36	58.71	38.31	74.13	0.2186	0.4095
Stable terminal interval	83.10	98.59	85.21	99.30	0.0392	0.0640

5.5. Identification Capability Analysis for High-Collapse-Pressure Intervals

It should be noted that the high-collapse-pressure intervals analyzed in this subsection are mainly located above the predefined target interval of 7900.000–7994.100 m. Therefore, this analysis is regarded as a supplementary evaluation of the model's response to high-value anomalies outside the primary target interval.

From the perspective of engineering application, the key to collapse pressure prediction lies not only in reducing the average error, but also in correctly identifying high-collapse-pressure intervals. Based on the distribution characteristics of the collapse pressure curve in this well, depth points with collapse pressure equivalent density not lower than 1.70 g/cm³ are defined as high-collapse-pressure points in this study. The correspondence between the high-value intervals identified from the actual results and those predicted by the model is compared, as shown in Table 5. This threshold is used only for the statistical identification of high-value intervals in this section and does not represent a fixed design value for drilling fluid density.

As shown in Table 5, when the threshold is set to 1.70 g/cm³, the actual high-collapse-pressure intervals are mainly located at 7792.786–7819.800 m, 7862.300–7863.900 m, and 7875.200–7879.400 m. The high-value intervals predicted by the model are 7792.786–7819.500 m, 7862.100–7864.200 m, and 7875.400–7879.600 m, respectively. The three groups of high-value intervals generally correspond well in terms of depth position, indicating that the model can accurately identify intervals with relatively high wellbore instability risk.

From the classification results of high-value points, the model achieves a precision of 97.01%, a recall of 98.48%, and an F1-score of 97.74%. This indicates that the model not

only maintains good performance in terms of overall error metrics, but also achieves high consistency in identifying high-collapse-pressure intervals, which is more important from an engineering perspective. For drilling fluid density design, the identification results of such high-value intervals can help determine in advance the wellbore stability risk intervals that require particular attention.

Table 5. Identification results of high-collapse-pressure intervals with a threshold of 1.70 g/cm³.

Item	Actual High-Value Interval/m	Predicted High-Value Interval/m	Comparison Result
High-value interval 1	7792.786~7819.800	7792.786~7819.500	Generally consistent
High-value interval 2	7862.300~7863.900	7862.100~7864.200	Generally consistent
High-value interval 3	7875.200~7879.400	7875.400~7879.600	Generally consistent
Overall identification metrics	330 actual high-value points	335 predicted high-value points	Precision = 97.01%, Recall = 98.48%, F1 = 97.74%

5.6. Engineering Applicability and Limitations

Based on the profile comparison, evaluation metrics, error distribution, and identification results of high-value intervals, the optimized model can effectively convert synthetic seismic attributes into continuous collapse pressure equivalent density profiles. The model shows relatively small errors in stable intervals and maintains a high identification rate in high-collapse-pressure intervals, indicating that this method can provide a reference for pre-drilling wellbore stability evaluation and drilling fluid density design in the target interval.

Meanwhile, the model prediction results still have certain limitations. First, the errors in high-value perturbation intervals are significantly larger than those in stable intervals, indicating that the model's response to local rapid variations still needs to be improved. Second, the validation in this study is mainly based on the comparison between predicted values and actual calculated results. If actual drilling responses such as borehole enlargement rate, electrical imaging logging, drilling drag, and pipe sticking can be further incorporated, the reliability of the model can be more effectively verified from the perspective of engineering phenomena. In addition, the current external validation is based on one actual well that was not involved in model training. Therefore, the present results demonstrate the feasibility of applying the proposed framework to an independent actual well, but they are not sufficient to fully characterize cross-well generalization. Finally, the reliability of the model is affected by the distribution of training samples, seismic wavelet extraction, the time–depth relationship, and the conversion of rock mechanical parameters. Therefore, when the model is subsequently extended to other blocks or lithological combinations, the training samples should be updated, and additional independent-well validation, including leave-one-well-out evaluation, should be conducted to further assess cross-well generalization. Future work should also incorporate uncertainty quantification to provide confidence intervals for the predicted collapse pressure and conduct model interpretability analyses to clarify the contributions of different input features. These improvements would enhance the transparency and reliability of the model and facilitate its practical deployment in different geological settings.

6. Conclusions

This study focuses on the prediction of collapse pressure equivalent density in the target interval of the study area and establishes a technical workflow consisting of virtual-well sample generation, synthetic seismic record construction, collapse pressure label calculation, and deep sequence model prediction. The main conclusions are as follows:

(1) Virtual-well samples were constructed based on the acoustic impedance distribution patterns of drilled wells. By combining actual seismic data for wavelet extraction, reflection coefficient calculation, and convolutional forward modeling, synthetic seismic

records were generated. This approach can form large-scale training data with geostatistical constraints and seismic response characteristics under the condition of limited sample size. The workflow maintains a good physical correspondence between the input synthetic seismic attributes and the output collapse pressure equivalent density, providing a basis for subsequent data-driven modeling.

(2) Under the unified training set, validation set, and evaluation metrics, the CNN-MultiLSTM model exhibits higher prediction accuracy than Random Forest, Polynomial Regression, and Support Vector Regression. This result indicates that collapse pressure equivalent density prediction depends not only on local seismic amplitude features, but is also controlled by continuous variations along the depth direction and the background trend of the interval. Therefore, deep learning models with convolutional local feature extraction capability and sequence memory capability are more suitable for this type of problem.

(3) On the basis of the CNN-MultiLSTM sequence-learning results, further structural optimization showed that simply increasing the number of training samples or expanding the local window provided only limited improvement, whereas introducing the normalized depth channel significantly enhanced the model's ability to represent positional constraints and the compaction background. The modeling strategy was therefore further extended from local-window point-wise prediction to whole-well sequence prediction, and residual dilated convolutions were introduced to enhance the representation of long-range depth-context information. With whole-well sequence input and whole-well curve output, the optimized model can make fuller use of both local waveform variations and long-range contextual information. The optimized whole-well sequence model achieves an MAE of 0.00262 g/cm^3 , an RMSE of 0.00352 g/cm^3 , a MAPE of 0.241% , and an R^2 of 0.9774 on the internal validation set, indicating that extending the modeling range from local windows to whole-well sequences is a key factor in improving the prediction accuracy of continuous collapse-pressure profiles.

(4) The external validation results using an actual well show that the model-predicted curve is in good agreement with the actual collapse pressure equivalent density curve in terms of the overall trend. For the full-well comparison interval, the MAE, RMSE, MAPE, and R^2 are 0.0246 g/cm^3 , 0.0410 g/cm^3 , 1.906% , and 0.9712 , respectively. For the target validation interval, the MAE, RMSE, MAPE, and R^2 are 0.0182 g/cm^3 , 0.0278 g/cm^3 , 1.640% , and 0.7356 , respectively. The errors are mainly concentrated in high-value perturbation intervals where the collapse pressure increases or decreases rapidly, whereas the prediction errors in stable intervals are relatively small. This indicates that the model has good tracking capability for stable backgrounds, but its response to local abrupt changes still has room for further improvement.

(5) As a supplementary analysis outside the primary target interval, the model was further evaluated for its capability to identify high-collapse-pressure intervals. When 1.70 g/cm^3 was used as the threshold for identifying high-collapse-pressure intervals, the predicted high-value intervals generally corresponded to the actual high-value intervals, with precision, recall, and F1-score values of 97.01% , 98.48% , and 97.74% , respectively. These results indicate that the model can capture high-value anomalies in the extended comparison interval and provide supplementary information for identifying key wellbore-stability risk intervals. Future studies still need to conduct independent validation using more actual drilling well data, borehole enlargement rates, electrical imaging logging, and drilling complications, and to recalibrate the seismic wavelet, time–depth relationship, and virtual-well sample distribution for different blocks.

Author Contributions: Conceptualization, N.L. and Y.X.; methodology, N.L., W.F. and Y.X.; software, J.L. and W.F.; validation, N.L., W.F. and Z.Z.; formal analysis, N.L., W.F. and Z.Z.; investigation, N.L., J.L. and Z.Z.; resources, Y.X. and Z.Z.; data curation, N.L., J.L., W.F. and Z.Z.; writing—original draft preparation, N.L. and Z.Z.; writing—review and editing, N.L., Y.X. and Z.Z.; visualization, J.L. and W.F.; supervision, Y.X.; project administration, Y.X.; funding acquisition, Y.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (Grant NO. 52374021, Grant NO. U24B2029).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: Author Ning Li and Wentong Fan are employed by the company PetroChina Tarim Oilfield Company. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Ibrahim, A. A review of mathematical modelling approaches to tackling wellbore instability in shale formations. *J. Nat. Gas Sci. Eng.* **2021**, *89*, 103870. [\[CrossRef\]](#)
2. Asaka, M.; Holt, R.M. Anisotropic wellbore stability analysis: Impact on failure prediction. *Rock Mech. Rock Eng.* **2021**, *54*, 583–605. [\[CrossRef\]](#)
3. Zhang, G.; He, S.; Tang, M.; Kong, L. The mechanism and countermeasures of inclined well wellbore instability in Dibe deep coal seam. *J. Pet. Explor. Prod. Technol.* **2022**, *12*, 2833–2848. [\[CrossRef\]](#)
4. Dahrabou, A.; Valley, B.; Meier, P.; Brunner, P.; Alcolea, A. A systematic methodology to calibrate wellbore failure models, estimate the in-situ stress tensor and evaluate wellbore cross-sectional geometry. *Int. J. Rock Mech. Min. Sci.* **2022**, *149*, 104935. [\[CrossRef\]](#)
5. Al-Ajmi, A.M.; Zimmerman, R.W. Stability analysis of vertical boreholes using the Mogi-Coulomb failure criterion. *Int. J. Rock Mech. Min. Sci.* **2006**, *43*, 1200–1211. [\[CrossRef\]](#)
6. Zhang, J. Borehole stability analysis accounting for anisotropies in drilling to weak bedding planes. *Int. J. Rock Mech. Min. Sci.* **2013**, *60*, 160–170. [\[CrossRef\]](#)
7. Chang, C.; Zoback, M.D.; Khaksar, A. Empirical relations between rock strength and physical properties in sedimentary rocks. *J. Pet. Sci. Eng.* **2006**, *51*, 223–237. [\[CrossRef\]](#)
8. Horsrud, P. Estimating mechanical properties of shale from empirical correlations. *SPE Drill. Complet.* **2001**, *16*, 68–73. [\[CrossRef\]](#)
9. Bowers, G.L. Pore pressure estimation from velocity data: Accounting for overpressure mechanisms besides undercompaction. *SPE Drill. Complet.* **1995**, *10*, 89–95. [\[CrossRef\]](#)
10. Zhang, J. Pore pressure prediction from well logs: Methods, modifications, and new approaches. *Earth-Sci. Rev.* **2011**, *108*, 50–63. [\[CrossRef\]](#)
11. Bergen, K.J.; Johnson, P.A.; de Hoop, M.V.; Beroza, G.C. Machine learning for data-driven discovery in solid Earth geoscience. *Science* **2019**, *363*, eaau0323. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195–204. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Dramsch, J.S. 70 years of machine learning in geoscience in review. *Adv. Geophys.* **2020**, *61*, 1–55. [\[CrossRef\]](#)
14. Karpatne, A.; Atluri, G.; Faghmous, J.H.; Steinbach, M.; Banerjee, A.; Ganguly, A.; Shekhar, S.; Samatova, N.; Kumar, V. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 2318–2331. [\[CrossRef\]](#)
15. Raissi, M.; Perdikaris, P.; Karniadakis, G.E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **2019**, *378*, 686–707. [\[CrossRef\]](#)
16. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
17. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [\[CrossRef\]](#)
18. Smola, A.J.; Scholkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [\[CrossRef\]](#)
19. Shan, L.; Liu, Y.; Tang, M.; Yang, M.; Bai, X. CNN-BiLSTM hybrid neural networks with attention mechanism for well log prediction. *J. Pet. Sci. Eng.* **2021**, *205*, 108838. [\[CrossRef\]](#)

20. Zeng, L.; Ren, W.; Shan, L.; Huo, F. Well logging prediction and uncertainty analysis based on recurrent neural network with attention mechanism and Bayesian theory. *J. Pet. Sci. Eng.* **2022**, *208*, 109458. [\[CrossRef\]](#)
21. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [\[CrossRef\]](#)
22. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [\[CrossRef\]](#)
23. Kiranyaz, S.; Avedi, O.; Abdeljaber, O.; Ince, T.; Gabbouj, M.; Inman, D.J. 1D convolutional neural networks and applications: A survey. *Mech. Syst. Signal Process.* **2021**, *151*, 107398. [\[CrossRef\]](#)
24. Fawaz, H.I.; Forestier, G.; Weber, J.; Idoumghar, L.; Muller, P.-A. Deep learning for time series classification: A review. *Data Min. Knowl. Discov.* **2019**, *33*, 917–963. [\[CrossRef\]](#)
25. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Lindemann, B.; Müller, T.; Vietz, H.; Jazdi, N.; Weyrich, M. A survey on long short-term memory networks for time series prediction. *Procedia CIRP* **2021**, *99*, 650–655. [\[CrossRef\]](#)
27. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
28. Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
29. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
30. Dutta, N.C. Geopressure prediction using seismic data: Current status and the road ahead. *Geophysics* **2002**, *67*, 2012–2041. [\[CrossRef\]](#)
31. Su, Y.; Cao, D.; Liu, S.; Hou, Z.; Feng, J. Seismic impedance inversion based on deep learning with geophysical constraints. *Geoenergy Sci. Eng.* **2023**, *225*, 211671. [\[CrossRef\]](#)
32. Meng, J.; Wang, S.; Niu, G.; Sang, W.; Geng, W.; Cheng, W. Seismic impedance inversion using a multi-input neural network with a two-step training strategy. *Geophys. Prospect.* **2024**, *72*, 107–124. [\[CrossRef\]](#)
33. Geng, Z.; Wang, Y. Physics-guided deep learning for predicting geological drilling risk of wellbore instability using seismic attributes data. *Eng. Geol.* **2020**, *279*, 105857. [\[CrossRef\]](#)
34. The MathWorks Inc. *MATLAB*, Version R2022b; The MathWorks Inc.: Natick, MA, USA, 2022.
35. Cai, R.; Sun, C.; Yao, Z.; Li, S. A depth-variant seismic wavelet extraction method for basis pursuit inversion with an impedance trend constraint. *Geophysics* **2024**, *89*, R275–R286. [\[CrossRef\]](#)
36. Cooley, J.W.; Tukey, J.W. An algorithm for the machine calculation of complex Fourier series. *Math. Comput.* **1965**, *19*, 297–301. [\[CrossRef\]](#)
37. Gardner, G.H.F.; Gardner, L.W.; Gregory, A.R. Formation velocity and density—The diagnostic basics for stratigraphic traps. *Geophysics* **1974**, *39*, 770–780. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.