Springer Series in Geomechanics and Geoengineering

Jia'en Lin *Editor*

# Proceedings of the International Field Exploration and Development Conference 2023

Vol. 10

Springer

# Springer Series in Geomechanics and Geoengineering

Series Editor

Wei Wu, *Universität für Bodenkultur, Vienna, Austria*

Geomechanics deals with the application of the principle of mechanics to geomaterials including experimental, analytical and numerical investigations into the mechanical, physical, hydraulic and thermal properties of geomaterials as multiphase media. Geoengineering covers a wide range of engineering disciplines related to geomaterials from traditional to emerging areas.

The objective of the book series is to publish monographs, handbooks, workshop proceedings and textbooks. The book series is intended to cover both the state-of-the-art and the recent developments in geomechanics and geoengineering. Besides researchers, the series provides valuable references for engineering practitioners and graduate students.

Indexed by SCOPUS, EI Compendex, INSPEC, SCImago.

Jia’en Lin
Editor

# Proceedings of the International Field Exploration and Development Conference 2023

Vol. 10

Springer

*Editor*
Jia'en Lin
College of Petroleum Engineering
Xi'an Shiyou University
Xi'an, Shaanxi, China

ISSN 1866-8755 ISSN 1866-8763 (electronic)
Springer Series in Geomechanics and Geoengineering
ISBN 978-981-97-0271-8 ISBN 978-981-97-0272-5 (eBook)
https://doi.org/10.1007/978-981-97-0272-5

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Paper in this product is recyclable.

# Contents

**Artificial Intelligence and Big Data Application in Oil and Gas Fields**

# Artificial Intelligence and Big Data Application in Oil and Gas Fields

# Application of Remote Sensing Intelligent Monitoring Technology for Oil and Gas Well Exit and Ecological Restoration

Hong-ying Zhou[1,2](✉), Yu-kun Guo[3], Qian Ye[2], Yuan-long Li[4], and Zhi-guo Ma[1]

[1] Research Institute of Petroleum Exploration and Development, PetroChina, Beijing, China
zhouhys@petrochina.com.cn
[2] Beijing Normal University, Beijing, China
[3] China University of Geosciences Beijing, Beijing, China
[4] PetroChina International Co., Ltd., Beijing, China

**Abstract.** Green Sustainable Development of Oil and Gas Fields considers oil and gas exploration as well as environmental conservation. Currently, national ecological protection requires that oil and gas exploration wells be shut down and closed when they withdraw, and that oil and gas facilities be demolished to restore the surrounding geomorphology and ecology. As a result, the condition of ecological restoration of oil and gas sites withdrawal is an essential component to evaluate the ecological protection of oil and gas fields. In this article, multi-temporal high-resolution satellite remote sensing big data is employed to achieve intelligent monitoring and assessment of green recovery at oil and gas sites. The technical process of remote sensing intelligent monitoring of oil and gas well withdrawal and ecological restoration includes three steps: 1. Determine the different types of well sites. Identify well sites where oil and gas facilities depart using high-resolution remote sensing oil and gas well site interpretation markings; 2. Detect well site modification. 3. Evaluate well site recovery by using the GRNDVI vegetation growth index into the well site vegetation change over time. To evaluate well site recovery, set a threshold value based on change detection data. Using Dabusu in Jilin Oilfield and Liaohekou in Liaohe Oilfield as experimental areas, remote sensing monitoring results show that 18 well sites in Dabusu experimental region were withdrawn in 2019 and all achieved vegetation restoration; 15 well

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20–22 September 2023.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 3–20, 2024.
https://doi.org/10.1007/978-981-97-0272-5_1

sites in Liaohekou experimental region were withdrawn from April to November 2018 to achieve natural restoration; and PetroChina has achieved results in ecological restoration of oil and gas well site withdrawal; remote sensing intelligent monitoring technology of oil and gas well site withdrawal and ecological restoration can realize large-scale and large-quantity well site withdrawal as well as efficient and accurate vegetation restoration monitoring. This technology ought to be used and popularized.

**Keywords:** Ecological restoration · Oil and gas well sites · Remote sensing monitoring · Change detection · GRNDVI

## 1 Introduction

In recent years, the state has accelerated the promotion of ecological environment restoration and management, as well as the construction of green mines, promoted the green and sustainable development of energy and resource exploitation, practiced the development concept of "Clear waters and green mountains are as good as mountains of gold and silver," and required energy enterprises to consider environmental protection while exploring and developing, in order to achieve the goals of the Kyoto Protocol. When oil and gas exploration and development wells exit, they must shut down the wells, dismantle the oil and gas facilities, avoid and reduce environmental pollution and land damage in the mining area, and carry out landform restoration and ecological restoration of the surrounding environment, according to the requirements of ecological environment protection. As a result, the ecological restoration condition of oil and gas well exit is a significant component of assessing the environmental protection of oil and gas fields.

Currently, ecological restoration monitoring in the exit region of oil and gas well sites is mostly accomplished by recurrent field inspection, which is unsuitable for large-scale monitoring due to its high cost and lack of timeliness. Remote sensing technology provides the advantages of long-range monitoring, short return periods, cheap cost, and the ability to do large-scale vegetation restoration monitoring and evaluation on a regular basis. Because oil and gas well sites are small in scale, large in number, and widely distributed, and their remote sensing image features differ by region and industry, it is necessary to develop a set of oil and gas well exit and ecological restoration index parameters, as well as a remote sensing intelligent monitoring technology process.

## 2 Remote Sensing Data and Evaluation Indicators

### 2.1 Introduction to Remote Sensing Data

GF2 is China's first self-developed submeter civil optical remote sensing satellite, which was successfully launched on August 19, 2014, from China's Taiyuan Satellite Launch Center. The satellite has two 1 m panchromatic/4 m multispectral cameras with a 45 km imaging width. The revisit duration can be reduced to 5 days with the use of a side swing (see Table 1).

Panchromatic and multispectral pictures are among the GF2 data. Panchromatic pictures have a spectral range of 450 nm to 900 nm with a spatial resolution of around 1m.The blue light band has a spectral range of 450 nm–520 nm, the green light band has a spectral range of 520 nm–590 nm, the red light band has a spectral range of 630 nm–690 nm, and the near-infrared band has a spectral range of 770 nm–890 nm. It has a spatial resolution of around 4 m.

**Table 1.** GF2 satellite payload technical indicators

| load | Spectral segment number | Spectrum range (um) | Spatial resolution (m) | Width (km) | Side sway capacity | Time of revisit (days) |
|---|---|---|---|---|---|---|
| Panchromatic multispectral camera | 1 | 0.45 ~ 0.90 | 1 | 45 (combination of 2 cameras) | ± 35 | 5 |
| | 2 | 0.45 ~ 0.52 | 4 | | | |
| | 3 | 0.52 ~ 0.59 | | | | |
| | 4 | 0.63 ~ 0.69 | | | | |
| | 5 | 0.77 ~ 0.89 | | | | |

## 2.2 Construction of Evaluation Indicators

The vegetation restoration evaluation index utilized in this article is the growth root normalized differential vegetation index (GRNDVI), which is based on the classic normalized vegetation index (NDVI) and has been enhanced to better the description of vegetation growth.

The classic NDVI indicator is made up of a red light band and a near-infrared band that can rise with vegetation development, although it is strongly influenced by high vegetation covering, atmosphere, and soil backdrop. When the vegetation coverage rate is low, the NIR/R is small, and the NDVI is obviously affected by the soil background, resulting in a larger NDVI value than is actually the case; when the vegetation coverage rate is high, the NIR/R is large, and the vegetation's absorption of R is gradually saturated, but the reflection of NIR continues to increase, causing its value to change more slowly, and the NDVI value is not sensitive to the response of vegetation changes.

$$\mathrm{NDVI} = \frac{\mathrm{NIR} - \mathrm{R}}{\mathrm{NIR} + \mathrm{R}} \tag{1}$$

Because the NDVI value is larger than the actual situation when the NIR/R is small, and the NDVI response is not sensitive enough when the NIR/R is large, multiplying the two can combine the characteristics of the two to some extent to alleviate the shortcomings of NDVI; however, the form of the product leads to the strengthening of the high value part and the weakening of the low value part. To address this issue, we square the product result; at the same time, the NDVI value is theoretically between (−1,1). To prevent negative values, multiply the NDVI by 1.

$$\mathrm{GRNDVI} = \left[\left(\frac{\mathrm{NIR}}{\mathrm{R}}\right) * \left(\frac{\mathrm{NIR} - \mathrm{R}}{\mathrm{NIR} + \mathrm{R}} + 1\right)\right]^{0.5} \tag{2}$$

NIR refers to the surface reflectance in the near-infrared band, and R refers to the surface reflectance in the red band.

## 3 Technical Process

The technological process of remote sensing monitoring for oil and gas well withdrawal and ecological restoration is divided into four stages: data acquisition and pretreatment, establishment of interpretation marks and interpretation, establishment and calculation of vegetation growth indicators, change detection analysis, and restoration effect evaluation (see Fig. 1).

The image data is preprocessed to provide surface reflectance data using distinct time phase high-resolution remote sensing pictures (including red light and near-infrared bands), and its geometric correctness fulfills the criteria of change detection. The interpretation markings of well sites and distinct oil and gas facilities are established based on the preprocessed pictures, and the well sites where oil and gas facilities depart are recognized based on this. The GRNDVI vegetation growth index is introduced to quantitatively describe plant growth in the well site's exit area. The change in plant growth state in different times is assessed using the GRNDVI of distinct time phases, and the vegetation restoration of the well site where the oil and gas facilities are extracted is evaluated.

### 3.1 Data Pre-processing

Data pre-processing primarily consists of ortho-correction, radiometric correction, atmospheric correction, geometric registration, picture fusion, and other similar operations. During imaging, ortho rectification may minimize geometric distortion caused by terrain variation, sensor side view angle, attitude, and azimuth. Image fusion may take into account both the spatial resolution of a panchromatic image and the spectral resolution of a multispectral image, which is used to extract color and texture information for interpretation marks. Radiation correction changes the DN value to apparent reflectance, whereas atmospheric adjustment removes the impact of the atmosphere to convert apparent reflectance to surface reflectance, which is used to construct the GRNDVI vegetation growth index. Geometric registration guarantees that the spatial coordinates of several temporal pictures are constant, and it offers a correct data foundation for detecting and analyzing changes between various phases at different times.

### 3.2 Establishment of Interpretation Marks

Oil and gas well sites are distinguished by their modest scale, huge quantity, and widespread dispersion within the context of oilfield exploration and mining rights. The

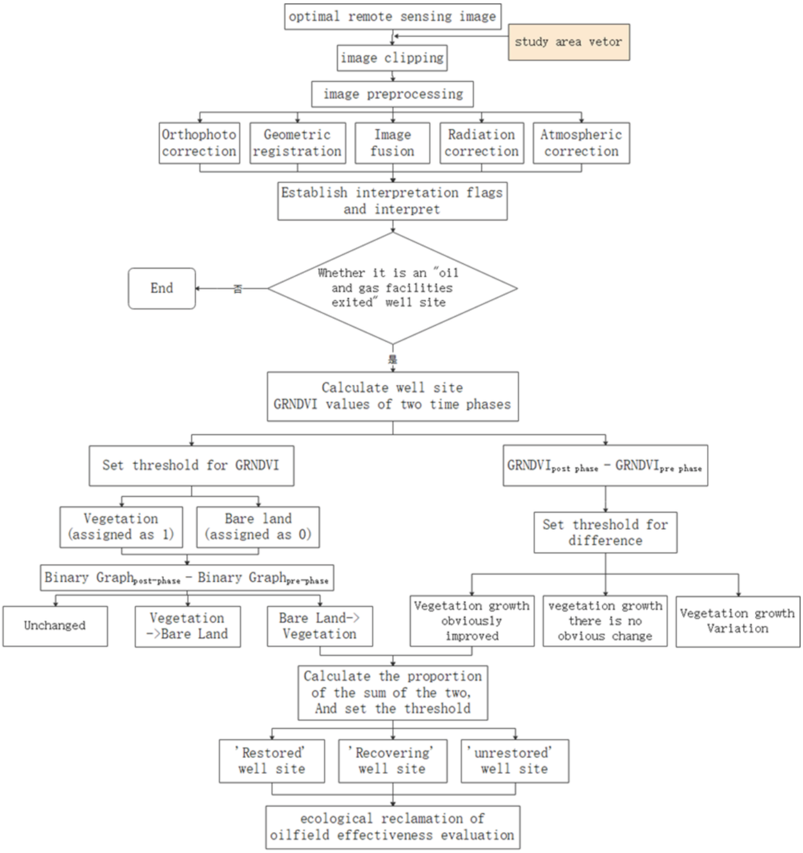

optimal remote sensing image
study area vetor
image clipping
image preprocessing
Orthophoto correction
Geometric registration
Image fusion
Radiation correction
Atmospheric correction
Establish interpretation flags and interpret
Whether it is an "oil and gas facilities exited" well site
否
End
是
Calculate well site GRNDVI values of two time phases
Set threshold for GRNDVI
Vegetation (assigned as 1)
Bare land (assigned as 0)
Binary Graph$_{post\text{-}phase}$ - Binary Graph$_{pre\text{-}phase}$
Unchanged
Vegetation ->Bare Land
Bare Land-> Vegetation
GRNDVI$_{post\ phase}$ - GRNDVI$_{pre\ phase}$
Set threshold for difference
Vegetation growth obviously improved
vegetation growth there is no obvious change
Vegetation growth Variation
Calculate the proportion of the sum of the two, And set the threshold
'Restored' well site
'Recovering' well site
'unrestored' well site
ecological reclamation of oilfield effectiveness evaluation

**Fig. 1.** Technical flow chart

remote sensing interpretation indications are summarized in Table 3 based on the color, geometric form, texture, azimuth relationship, size, and other features of oil and gas well sites and oil and gas facilities on high-resolution remote sensing pictures.

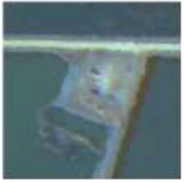
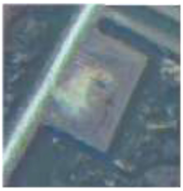
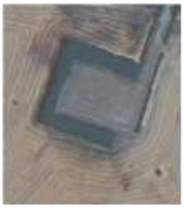
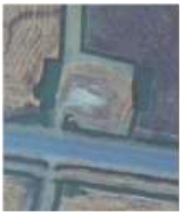
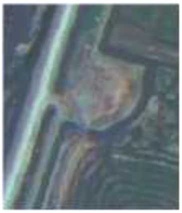
### 3.3 Change Detection

Change detection is comprised of two components: qualitative classification of change kinds and quantitative assessment of change degree.

Qualitative categorization of well site change types: based on the preprocessed picture, the vegetation growth index GRNDVI is extracted, and the well site area is classified into "vegetation" and "bare land" by setting threshold values of 1 and 0, respectively. Then, using varying time differences, three change types were obtained: "unchanged", "vegetation -> bare land", and "bare land -> vegetation".

Quantitative evaluation of well site change degree: based on the preprocessed picture, the vegetation growth index GRNDVI is retrieved, and quantitative change data is derived by comparing GRNDVI values in different time phases. Threshold values are established

**Table 2.** Remote sensing interpretation marks of different types of well sites

| Type | GF2 image | Description | Field photos |
|---|---|---|---|
| (Well site in use) Production well | <br> | Oil and gas well sites are generally square bare ground with light color and rectangular shape;Wired roads are connected with it and interconnected with other well sites through roads,;There are different types of pumping units in the well site of the oil production well in use, and the common beam pumping units, due to the difference in imaging angles, have a "sickle" or "long strip" shape, and the pixels in the front of the strip are blurred;Under different lighting conditions, dark shadows like "straight lines", "broken lines" or "lumps" will appear near some pumping units. | <br>7<br> |
| (Exited from the well site) Completely naked |  | The shape of abandoned wells is mostly rectangular, occasionally with other irregular shapes;Its hue is mostly light, which is obviously different from the surrounding background;There are generally no oil and gas facilities inside, and sometimes a small amount of vegetation can be seen | |
| (Well site under restoration) Seminaked |  | Most of the well sites under restoration are rectangular in shape;There is no obvious difference between the restored vegetation area and the surrounding background tone;Generally, there are no oil and gas facilities inside, but sometimes there are cement bases for pumping units. The bases are mostly light colored rectangles, which are obviously different from the surrounding background. | |
| Well site restore |  | The tone and texture of the recovered well site are completely consistent with the surrounding background;The boundaries of a few well sites are indistinct and visible, and most of them have no fixed shape after full recovery, so it is almost impossible to distinguish them from the surrounding background. | |

for evaluating quantitative change data, and the modified well site regions are classified as “growing better,” “growing worse,” or “growing without significant change.”

Calculate the proportion of “bare land -> vegetation” and “obviously growing better” at the well site, and establish a threshold to split the well site into “recovered”, “recovering (good recovery),” and “not recovered (general recovery).” When computing the percentage, the overlapping section of the two types of regions should be deleted to avoid remeasurement.

# 4 Method Experiment

According to the survey, Liaohe Oilfield gradually withdrawn oil wells from the nature reserve from 2013 to 2019, and insisted on restoring the coverage of vegetation such as reeds primarily through natural recovery supplemented by manual recovery; Jiaohe Oilfield completed the withdrawal of oil wells from the Dabusu Nature Reserve in 2018, and carried out the ecological reclamation of the withdrawn oil wells. In this paper, the demonstration areas of Liaohekou and Dabusu are chosen to conduct remote sensing intelligent monitoring technology and method experiments for oil and gas well withdrawal and ecological restoration, as well as to further evaluate the effectiveness of Liaohe Oilfield and Jilin Oilfield in ecological reclamation.

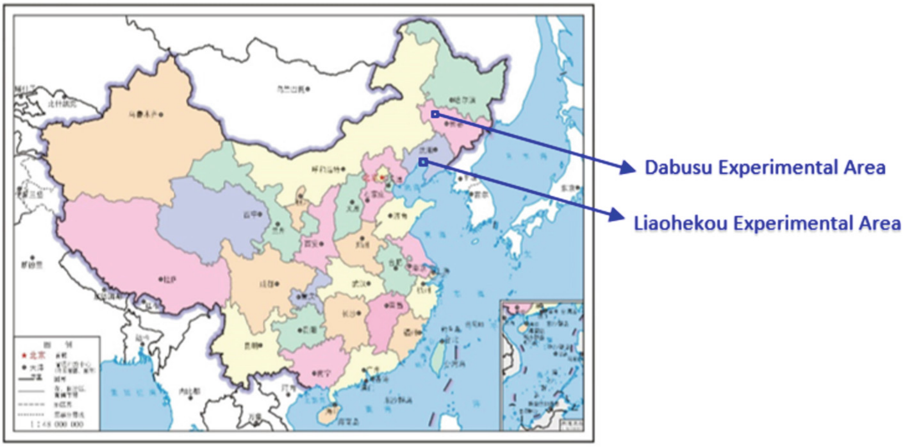


**Fig. 2.** Location of Demonstration Area

## 4.1 Dabusu Experimental Region

Dabusu Lake is located in the south of Songnen Plain, with the geographical coordinates of 123.60°–123.71° E and 44.75°–44.84° N. It was formed in the late Pleistocene, and its water quality is strongly alkaline, with a PH of 10–11. The average temperature in January in this area is about −14.5 °C, which is the lowest temperature in the whole year; The average temperature in July is about 25 °C, the highest temperature in the whole year; The average annual rainfall is about 400 mm, of which the proportion of rainfall in June, July and August accounts for more than 70% of the total rainfall of the year. The experimental area is located on the northeast bank of the Great Busu Lake.

### 4.1.1 Introduction to Remote Sensing Data

The remote sensing images of GF2 satellite were screened according to the three principles of 2017–2019, the same period in summer and less than 5% cloud cover, and the images of August 25, 2017 and September 4, 2019 were selected as the remote sensing data of the study area (see Fig. 3 and Fig. 4).

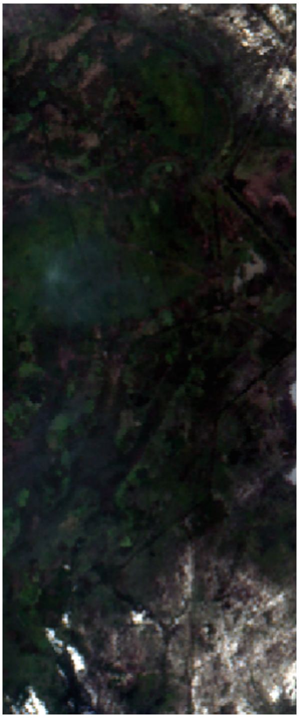


**Fig. 3.** Image of Dabusu on August 30, 2017

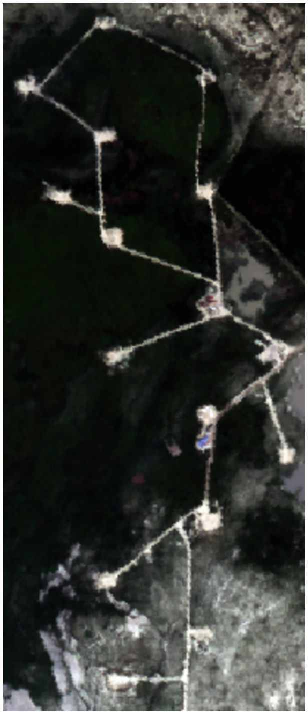


**Fig. 4.** Image of Dabusu on September 4, 2019

### 4.1.2 Change Detection Experiment

The experiment procedure is divided into three stages: theme information extraction of the well site and GRNDVI, multi temporal remote sensing image change detection, and well site vegetation restoration evaluation.

1. **Extraction of thematic information**

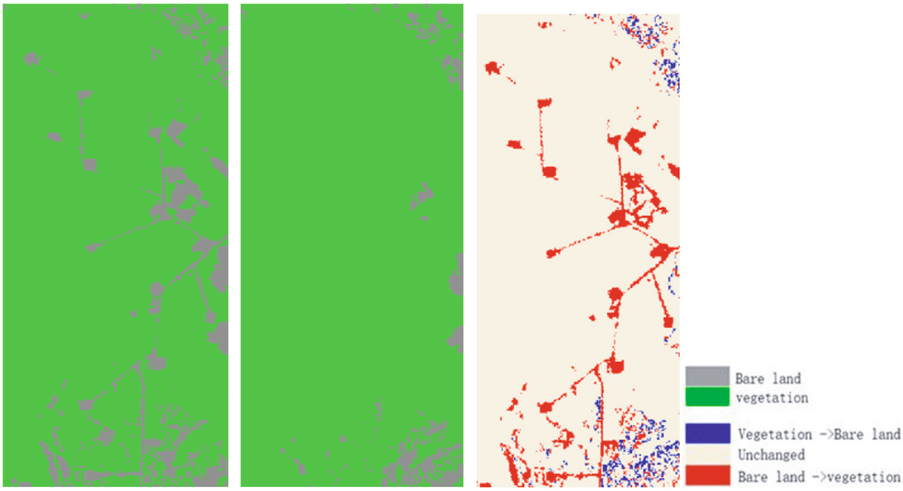
The two time phases' remote sensing pictures are preprocessed with radiometric correction, image fusion, and other techniques, and oil and gas wells are detected and retrieved using the established remote sensing interpretation markers. At the same time, the GRNDVI vegetation growth index for each time phase is computed.

2. **Detection of Changes**

The detection of change is separated into two parts: qualitative classification of change kinds and quantitative evaluation of change degree.

The first is the identification of change kinds by qualitative categorization. The GRNDVI criteria for the two stages are established. In the Dabusu experimental region, the GRNDVI threshold of the remote sensing picture is set to 1.4. The well site area is classified into "vegetation" and "bare land" based on the threshold segmentation findings, with values of 0 and 1, respectively. The categorization results were divided into two time periods, and three forms of change were detected: "unchanged" "vegetation -> bare land" and "bare land -> vegetation" (see Fig. 5). Choose the "bare land -> vegetation" type region, which is referred to as the "vegetation restoration area".

The second step is a quantitative assessment and study of the degree of change. The GRNDVI index of the two time phases is distinguished, and the difference result's threshold value is established. Following the experiment, the threshold is set to 0.2, and the values are allocated to -1, 0, and 1. The well site is rated as "growing better", "growing worse" or "growing with no significant change" (see Fig. 6). Choose the "growth recovery area", which has a rating of "growing better".



(a) 2017 GRNDVI map (b) 2019GRNDVI map (c) bare land ->vegetation

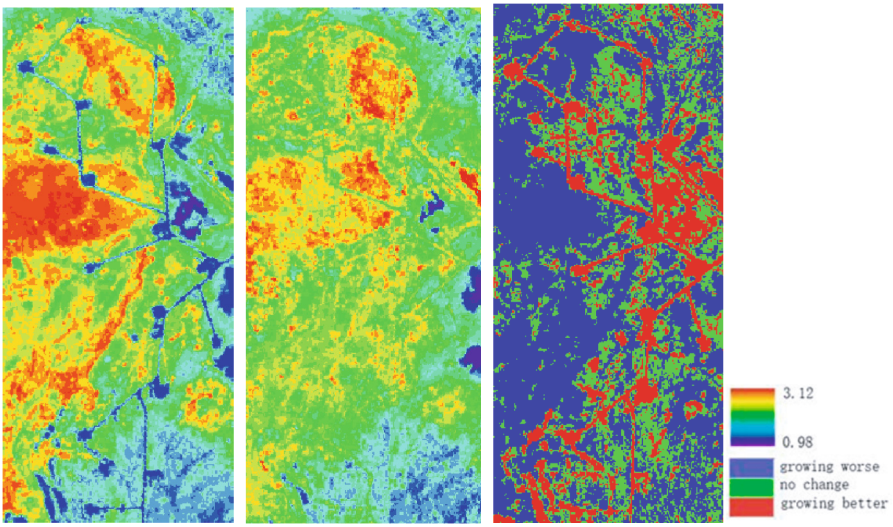
**Fig. 5.** Qualitative detection results of change types in Dabusu region

3. **Recovery Evaluation**

Count the "vegetation restoration area" and "growth restoration area" within the well site, calculate the proportion of the combined area of the two areas in the total area of the well site, and set a threshold for this proportion to further evaluate the well site's vegetation restoration. In conjunction with the unique conditions of the Dabusu region, the threshold value is established at 0.8, and the vegetation restoration of the well site is graded as zero levels of "under restoration" and "restored". Table 3 displays the statistical data.

### 4.1.3 Analysis of Experimental Results

All oil wells in the Dabusu experimental region have completed the withdrawal of oil and gas facilities, according to the evaluation results of vegetation restoration in Table 1. According to the statistical data, each well site has two phenomena. Firstly, the "number of vegetation restoration pixels" and "number of overlapping pixels" of all well sites are very close, and in some cases equal, indicating that most areas in the well site have been



(a) 2017 GRNDVI map (b) 2019GRNDVI map (c) GRNDVI change map

**Fig. 6.** Quantitative evaluation results of change degree in Dabusu region

restored from "bare land" to "vegetation" and their growth has also improved significantly. Secondly, the "number of growth restoration pixels" of each well site exceeds the "number of vegetation restoration pixels," indicating that there was a significant amount of vegetation in some areas of each well site prior to the withdrawal of oil and gas facilities, but there is still much room for growth improvement. Finally, except for Well Site 1 and Well Site 4, where the recovery rate is around 72.4% and 75.6%, the recovery rate of the remaining well sites is greater than 80%, implying that there are 16 "recovered" well sites in the Dabusu experimental region.

## 4.2 Liaohekou Experimental Region

The Liaohe River estuary is located in Panjin City, Liaoning Province, at the mouth of the Liaohe River. Its geographical coordinates are 121.86°–121.88° E and 41.038°–41.045° N, with an average altitude of about 2 m. The region belongs to temperate semi humid monsoon climate, with four distinct seasons. The annual average temperature is about 8.5 °C, of which the average temperature in summer is about 23 °C, and the average temperature in winter is about - 8 °C; The average annual rainfall is about 650mm, of which heavy rainfall is easy to occur in June, July and August. The Liaohe River estuary experiment is divided into two parts, north and south.

### 4.2.1 Introduction to Remote Sensing Data

In order to monitor the vegetation restoration in Liaohekou experimental region, GF2 satellite remote sensing images with cloud cover less than 5% in 2017–2022 and the same period in summer were screened, and the images on August 30, 2017 and August

**Table 3.** Results of Dabusu experiment and evaluation of vegetation restoration

| Well Site No | facilities Exit evaluation | Analysis of well site green restoration pixels | | | | | | Vegetation restoration evaluation |
|---|---|---|---|---|---|---|---|---|
| | | vegetation restoration pixels | growth recovery pixels | coincident pixels | Restore pixels merging | pixels in the well site | Percentage of recovered pixels | |
| 1 | Exited | 26 | 49 | 25 | 50 | 69 | 0.724638 | Recovering |
| 2 | Exited | 81 | 116 | 81 | 116 | 135 | 0.859259 | Recovered |
| 3 | Exited | 94 | 124 | 94 | 124 | 149 | 0.832215 | Recovered |
| 4 | Exited | 61 | 90 | 55 | 96 | 127 | 0.755906 | Recovered |
| 5 | Exited | 67 | 101 | 66 | 102 | 118 | 0.864407 | Recovered |
| 6 | Exited | 173 | 225 | 173 | 225 | 244 | 0.922131 | Recovered |
| 7 | Exited | 52 | 78 | 52 | 78 | 96 | 0.8125 | Recovered |
| 8 | Exited | 135 | 157 | 135 | 157 | 161 | 0.975155 | Recovered |
| 9 | Exited | 68 | 107 | 68 | 107 | 122 | 0.877049 | Recovered |
| 10 | Exited | 25 | 58 | 25 | 58 | 69 | 0.84058 | Recovered |
| 11 | Exited | 157 | 198 | 159 | 196 | 219 | 0.894977 | Recovered |
| 12 | Exited | 66 | 88 | 66 | 88 | 107 | 0.82243 | Recovered |
| 13 | Exited | 118 | 159 | 118 | 159 | 180 | 0.883333 | Recovered |
| 14 | Exited | 84 | 143 | 84 | 143 | 176 | 0.8125 | Recovered |
| 15 | Exited | 90 | 97 | 77 | 110 | 136 | 0.808824 | Recovered |
| 16 | Exited | 89 | 83 | 73 | 99 | 119 | 0.831933 | Recovered |
| 17 | Exited | 53 | 52 | 51 | 54 | 64 | 0.84375 | Recovered |
| 18 | Exited | 103 | 101 | 100 | 104 | 108 | 0.962963 | Recovered |

14, 2020 were selected as the basic data of Liaohekou experimental south region (see Fig. 7), The images on August 30, 2017 and July 14, 2022 are selected as the research data of Liaohekou experimental north region (see Fig. 8).

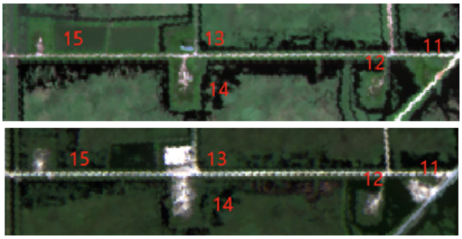


**Fig. 7.** Image of Liaohekou south region (left: August 30, 2017; right: August 14, 2020)

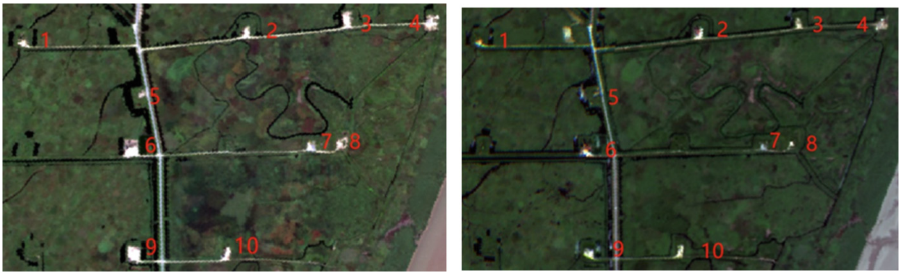


**Fig. 8.** Image of Liaohekou north region (left: August 30, 2017; right: July 14, 2022)

### 4.2.2 Change Detection Experiment

1. **Extraction of thematic information**

The remote sensing images of the two time phases are preprocessed by radiometric correction, image fusion, etc., and oil and gas wells are identified and extracted according to the established remote sensing interpretation marks. At the same time, the GRNDVI vegetation growth index of the two time phases is calculated respectively.

2. **Detection of changes**

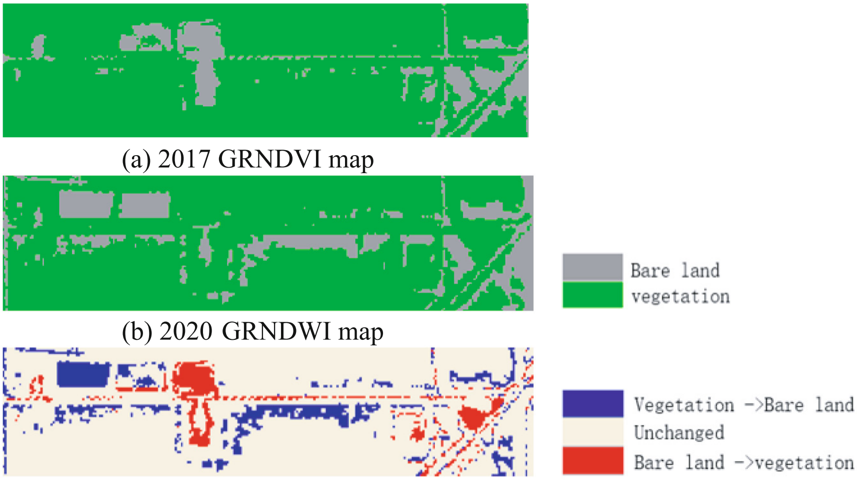
The detection of change is separated into two parts: qualitative classification of change kinds and quantitative evaluation of change degree. The procedure is the same as in the Dabusu experimental region. The threshold values for change detection fluctuate due to the geographical environment and climate features of the Liaohekou region, as well as the effect of the time equivalent components of the photos utilized.

The first is the identification of change kinds by qualitative categorization. The GRNDVI threshold values in the south and north portions of the experiment are 1.6 and 1.4, respectively. The well site area is classified into "vegetation" and "bare land" based on the threshold segmentation findings, with values of 0 and 1, respectively.The classification results were divided into two time periods, and three types of changes were detected: "unchanged", "vegetation -> bare land" and "bare land -> vegetation" (see Fig. 9 and Fig. 11). Choose the "bare land -> vegetation" type region, which is referred to as the "vegetation restoration area".

The second step is a quantitative assessment and study of the degree of change. The GRNDVI index of the two time phases of the southern and northern experimental regions is distinguished, and the difference result's threshold value is determined. The southern region of Liaohekou estuary's threshold value is set to 0.3, while the northern region of Liaohe estuary's threshold value is set to 0.2, with the values assigned to $-1$, 0, and 1. The well site is graded as "growing better", "growing worse" or "growing without significant change" (see Fig. 10 and Fig. 12). Choose the "growth recovery area", which has a rating of "growing better".

3. **Recovery Evaluation**

Count the "vegetation restoration area" and "growth restoration area" within the well site, calculate the proportion of the combined area of the two areas in the total

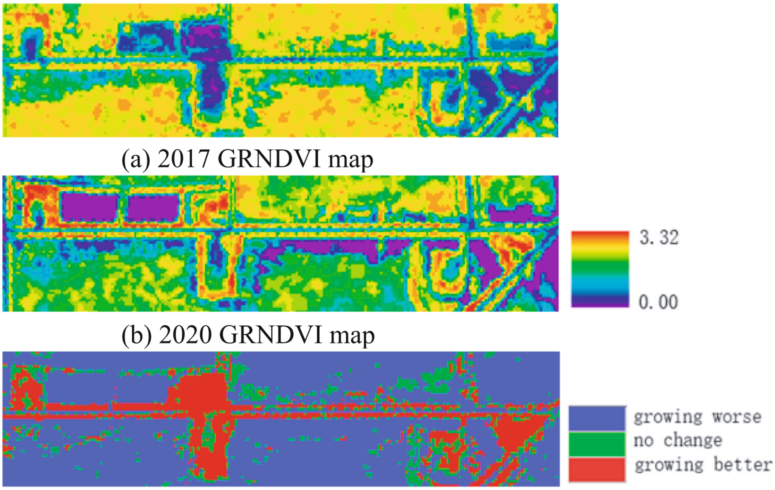


(a) 2017 GRNDVI map

(b) 2020 GRNDWI map

(c) Bare land->vegetation map

**Fig. 9.** Qualitative detection results of change types in Liaohekou south region



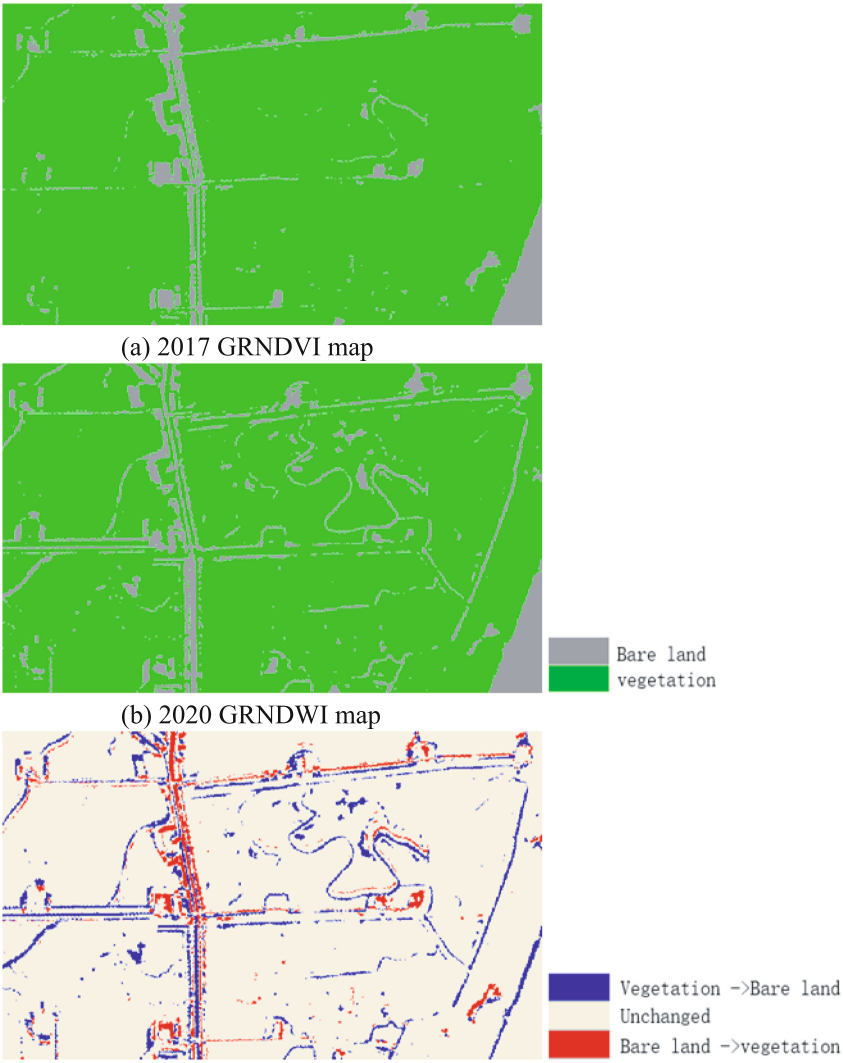
(a) 2017 GRNDVI map

(b) 2020 GRNDVI map

(c) GRNDVI change map

**Fig. 10.** Quantitative evaluation results of change degree in Liaohekou south region

area of the well site, and set a threshold for this proportion to further evaluate the well site's vegetation restoration.In conjunction with the current situation of the Liaohekou experimental region, the threshold values are set to 0.5 and 0.8, and the vegetation restoration of the well site is evaluated as "average restoration", "good restoration" or "restored", with "average restoration" and "good restoration" being the status of vegetation restoration. Table 4 displays the statistical data.



(a) 2017 GRNDVI map

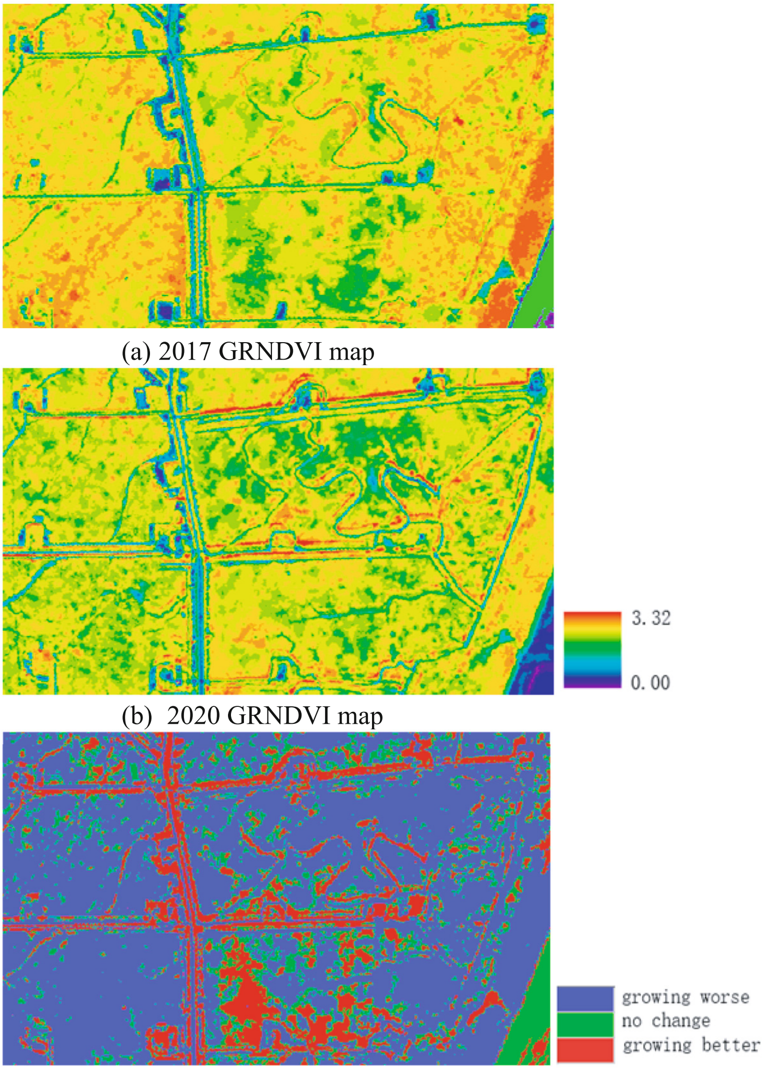
(b) 2020 GRNDWI map

(c) bare land->vegetation map

**Fig. 11.** Qualitative detection results of change types in Liaohekou north region

### 4.2.3 Analysis of Experimental Results

All oil wells in the Liaohekou experimental region have completed the withdrawal of oil and gas facilities, according to the evaluation findings of vegetation restoration in Table 2. Similar to the situation in the Dabusu experimental region, the "number of vegetation restoration pixels" and "number of overlapping pixels" of all well sites are very close, and the "number of growth restoration pixels" of all well sites is greater than the "number of vegetation restoration pixels," indicating that most areas are recovered from "bare land" to "vegetation" at the same time, according to the statistical data



(a) 2017 GRNDVI map

(b) 2020 GRNDVI map

(c) GRNDVI change map

**Fig. 12.** Quantitative evaluation results of change degree in Liaohe north region

in Table 2. Its growth has also improved dramatically, indicating that there was a big quantity of vegetation in some regions of each well site prior to the removal of oil and gas infrastructure, and its growth still has a lot of potential to develop. However, unlike the Dabusu experimental region, only five well sites have been found in the Liaohekou experimental region, namely well sites 7, 8, 9, 13, and 14. There are another eight well locations with strong recovery effects: 1, 3, 5, 6, 10, 11, 12, and 15. Finally, two well sites, well sites 2 and 4, had moderate recovery effects.

**Table 4.** Results of Liaohekou experiment and evaluation of vegetation restoration

| Well Site No | facilities Exit evaluation | Analysis of well site green restoration pixels | | | | | | Vegetation restoration evaluation |
|---|---|---|---|---|---|---|---|---|
| | | vegetation restoration pixels | growth recovery pixels | coincident pixels | Restore pixels merging | pixels in the well site | Percentage of recovered pixels | |
| 1 | Exited | 61 | 128 | 59 | 130 | 208 | 0.625 | good |
| 2 | Exited | 30 | 56 | 30 | 56 | 135 | 0.414815 | general |
| 3 | Exited | 72 | 163 | 68 | 167 | 238 | 0.701681 | good |
| 4 | Exited | 44 | 134 | 39 | 139 | 321 | 0.433022 | general |
| 5 | Exited | 43 | 52 | 41 | 54 | 74 | 0.72973 | good |
| 6 | Exited | 194 | 281 | 182 | 293 | 421 | 0.695962 | good |
| 7 | Exited | 69 | 124 | 69 | 124 | 140 | 0.885714 | Recovered |
| 8 | Exited | 145 | 185 | 149 | 181 | 197 | 0.918782 | Recovered |
| 9 | Exited | 149 | 230 | 141 | 238 | 288 | 0.826389 | Recovered |
| 10 | Exited | 56 | 97 | 56 | 97 | 151 | 0.642384 | good |
| 11 | Exited | 146 | 245 | 142 | 249 | 342 | 0.72807 | good |
| 12 | Exited | 39 | 82 | 31 | 90 | 133 | 0.676692 | good |
| 13 | Exited | 142 | 262 | 134 | 270 | 328 | 0.823171 | Recovered |
| 14 | Exited | 272 | 325 | 271 | 326 | 351 | 0.928775 | Recovered |
| 15 | Exited | 32 | 102 | 28 | 106 | 148 | 0.716216 | good |

## 5 Conclusion

Given the characteristics of small, large, and widely distributed oil and gas well sites, in order to carry out large-scale oil and gas well site exit and vegetation restoration monitoring efficiently, accurately, and intelligently, this paper proposes an intelligent remote sensing monitoring technology for oil and gas well exit and ecological restoration based on multi temporal high-resolution satellite remote sensing data, and chooses GF2 satellite remote sensing images. Using Liaohekou and Dabusu as experimental places, the technological procedure is used to monitor and analyze the removal of oil and gas well sites as well as vegetation restoration. The results show that remote sensing intelligent monitoring technology for oil and gas well withdrawal and ecological restoration is effective in monitoring and counting the vegetation restoration of oil and gas facility withdrawal well sites, and has low cost and high efficiency advantages that traditional methods do not have.

The experimental area's statistical data show that the vegetation restoration effect of the well sites with oil and gas facilities withdrawn in Liaohekou experimental region is good, compared to that of the well sites with oil and gas facilities withdrawn in Dabusu experimental region, the vegetation restoration effect of the well sites with oil and gas facilities withdrawn in Dabusu experimental region is more significant.

## References

1. Zhang, Z., Yang, J., Li, R., Li, Q.: Application of remote sensing technology in oilfield monitoring. Jilin Geol. **40**(03), 37–41 (2021)
2. Wu, Q.: Assessments of Ecological Environment in the central part of Huabei Oilfield based on RS and GIS. Jilin University (2022)
3. Chen, C.: Remote Sensing Assessments of Ecological Environment based on Game Theory in Daqing Oilfield. Jilin University (2021)
4. Wang, W.: Remote Sensing Monitoring and Evaluation of Mine Regreening Based on Time Series Feature Change.China University of Geosciences (Beijing), (2021)
5. Zhao, H.: Some Issues of Remote Sensing Based Crop Phenology and Condition Monitoring. WuhanUniversity (2010)
6. Zhao, H., Yang, Z., Li, L., Di, L.: Improvement and comparative analysis of indices of crop growth condition monitoring by remote sensing. J. Agricult. Eng. **27**(01), 243–249+394 (2011)
7. Zhao, Y., Yang, J., Zhang, Z.: Preliminary study on the remote sensing monitoring of the oil and gas resources: a case study on the ordos basin. mineral exploration **10**(12), 2980–2989 (2019)
8. Ren, Q., Yang, W., Wang, C., Wei, W., Qian, Y.: Review of remote sensing image change detection. Comput. Appl. **41**(08), 2294–2305 (2021)
9. Su, J.: Application of remote sensing image map in assisting oilfield surface planning and design. Oil and Gas Field Surface Engineering **42**(01), 12–15 (2023)
10. Zhao, Y., Yang, J., Sun, Y., Chen, D.: A remote sensing method for judging the cross-border mining of oil and gas mines. Remote Sens. Natural Resources **34**(02), 30–36 (2022)
11. Wang, Y., Li, L.: Remote sensing monitoring for the oil and gas platform in the South China Sea. China Geol. Surv. **8**(03), 58–63 (2021)
12. Lord, B.: Remote sensing techniques for onshore oil and gas exploration. Lead. Edge **36**(1), 24–32 (2017)
13. Qin, Q., Zhang, Z., Chen, L., Wang, N., Zhang, C.: Oil and gas reservoir exploration based on hyperspectral remote sensing and super-low-frequency electromagnetic detection. J. Appl. Remote Sens. **10**(1), (2016)
14. Arslan, N., Majidi Nezhad, M., Heydari, A., Astiaso Garcia, D., Sylaios, G.: A principal component analysis methodology of oil spill detection and monitoring using satellite remote sensing sensors. Remote Sens. **15**(5), 1460 (2023)

15. Zhao, D., Tan, B., Zhang, H., Deng, R.: Monitoring marine oil spills in hyperspectral and multispectral remote sensing data by the spectral gene extraction (sge) method. Sustainability **14**(20), 13696 (2022)
16. Zhu, H., Jia, G., Zhang, Q., Zhang, S., Lin, X., Shuai, Y.: Detecting offshore drilling rigs with multitemporal NDWI: a case study in the Caspian sea. Remote Sensing **13**(8), 1576 (2021)

# Research on Prediction of the Effects of Oil-Increasing Measures Driven by Data

Lu Yang[1], Kai Zhang[1](✉), Li- ming Zhang[1], Hua- qing Zhang[2], Xia Yan[1], Pi-yang Liu[3], and Jun Yao[1]

[1] School of Petroleum Engineering, China University of Petroleum (East China), Qingdao, China
zhangkai@upc.edu.cn

[2] College of Science, China University of Petroleum (East China), Qingdao, China

[3] Civil Engineering School, Qingdao University of Technology, Qingdao, China

**Abstract.** A large number of major oil fields in China have entered the late stages of development, and the decreasing production is increasingly unable to meet the continuously growing demand for energy. Therefore, it is crucial for oilfield production to accurately and rapidly predict the effects of production-increasing measures based on existing data. This paper comprehensively considers three types of data: geological static parameters, production dynamic parameters, and process parameters of measures. Advanced machine learning algorithms such as random forest (RF), support vector regression (SVR), and extreme gradient boosting (XGBoost) are separately used, together with data augmentation techniques and Bayesian optimization algorithms to construct the different enhancing production through measures prediction model. The best prediction model is optimized by comparing the scores of each model. The results of a comprehensive comparison of various models based on the mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R2) indicate that the model based on the extreme gradient boosting algorithm performs the best. The application of data augmentation and optimization algorithms significantly improves the model performance. The accuracy of predicting the oil production enhancement effect for a given measure can reach over 90%. Compared with traditional methods for predicting the effects of measures, this paper addresses the issues of long computational time in numerical simulations and difficulty in exploring the mechanism of

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20–22 September 2023.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 21–30, 2024.
https://doi.org/10.1007/978-981-97-0272-5_2

oil production enhancement measures in depth, and achieves a rapid and accurate prediction of the multidimensional effect of measures for increasing oil production. This paper employs machine learning algorithms to fully explore the relationship between three types of data and oil production enhancement effects, accurately predicting the effect of measures for increasing oil production. It provides a technical foundation for selecting reasonable measures to increase oil production in oilfields and has certain guiding significance for actual production.

## 1 Introduction

As an irreplaceable strategic resource, petroleum plays a critical role in a country's power and economic development. Maintaining and increasing oil production has always been an important energy goal for nations. However, due to China's rapid economic growth over the past few decades, the country's consumption of petroleum has been steadily increasing. Nevertheless, as most of China's major oil fields have entered the middle and late stages of development, issues such as increased water content and reservoir damage have resulted in a decline in production capacity that is no longer sufficient to meet current energy demands [1]. This has led to a severe dependence on foreign oil and gas resources, with China's external oil dependency exceeding 70% in 2020. Large-scale oil imports could easily subject China to geopolitical risks, significantly threatening the country's energy security [2]. Therefore, implementing reasonable measures to increase oil production is imperative for China to address its energy gap, stabilize domestic economic development, and alleviate its energy crisis. However, with a wide variety of measures available and varying effectiveness, accurately predicting the effectiveness of such measures is crucial for oil field production.

In the field of measure effectiveness prediction, conventional methods such as the water flooding characteristic curve method and the Weng cycle method have limitations in their applicability due to various assumptions and complex formulas [3]. Although numerical simulations have been attempted to predict the effectiveness of measures, their applicability is restricted due to the complex mechanisms of measures to increase oil production and the expensive computations involved [4]. Research on machine learning-based measure effectiveness prediction is still in the exploratory stage, with a primary focus on production forecasting. There has been limited consideration of including process parameters in the evaluation of measures, as the limited sample size of measure wells restricts research in this direction to primarily fracturing methods [5].

In recent years, the revolutionary development of artificial intelligence (AI) technology has attracted widespread attention from various industries due to its powerful generalization ability and rapid response speed [6]. The petroleum industry has also accumulated a large amount of historical data in production, and machine learning has shown great potential in the field of petroleum engineering [7]. As an alternative data-driven approach, machine learning can extract information from a large amount of historical data and construct regression or classification prediction models [8]. Many supervised machine learning methods, including linear regression, support vector machines, neural networks, etc., have been used to predict production decline, optimize water injection schemes, characterize reservoir permeability, and generate complex geological facies [9].

Based on the above content, in order to accurately and rapidly predict the effect of oil recovery measures, this paper proposes a data-driven approach for predicting the effect of oil recovery measures. Advanced machine learning algorithms, including Random Forest (RF), Support Vector Regression Machine (SVR), and Extreme Gradient Boosting (XGBoost), are used to explore the influence of three types of data on the effect of oil recovery measures, namely geological static parameters, production dynamic parameters, and process parameters. A prediction model is built, and data augmentation is employed to address the problem of insufficient samples, which improves the quality of the sample dataset. In the hyperparameter tuning stage of the model, the Bayesian optimization algorithm is introduced to solve the problem of difficult manual parameter tuning and further improve the model accuracy. After comparative experiments, the XGBoost algorithm-based oil recovery measure effect prediction model is selected, and the accuracy in the test set can reach over 90%.

## 2 Methodology

### 2.1 Feature Engineering

Feature engineering is the process of taking the raw input data and creating new features. To make the raw data more informative, it selects, extracts, and transforms meaningful features from the raw data. Feature engineering involve various techniques, including data cleaning, data normalization, data scaling, data augmentation, data encoding, dimensionality reduction, and feature selection. The source data for this study is the actual recorded data from the oil field, which has poor data quality. Therefore, feature engineering is crucial in processing the data. In addition to common data cleaning, normalization, and correlation analysis, this paper also employs the SMOTE oversampling technique as a data augmentation method.

**Synthetic Minority Over-Sampling Technique (SMOTE)**
SMOTE [10] is an approach to the construction of classifiers from imbalanced datasets is described. It is used to address the problem of imbalanced class distribution in data by synthesizing data through a combination of over-sampling the minority class and under-sampling the majority class [11]. The specific steps are as follows:

(1) For each sample x in the minority class, calculate its k-nearest neighbors to all samples in the minority class using Euclidean distance as the metric. The formula is:

$$d(s_l, s_k) = \sqrt{\sum_{j=1}^{m}\left(s_{lj} - s_{kj}\right)^2} \tag{1}$$

(2) Determine a sampling rate based on the imbalance ratio and set a sampling multiplier N. For each minority sample x, randomly select several samples from its k-nearest neighbors, denoted as $x_n$.

(3) For each randomly selected neighbor $x_n$, construct a new sample with the original sample according to the following formula.

$$x_{\text{new}} = x + \text{rand}(0, 1) \times (\tilde{x} - x) \tag{2}$$

## 2.2 Regression Prediction Algorithm

**Support Vector Regression (SVR)**

SVR [12] is a type of machine learning algorithm used for regression analysis. It is based on the Support Vector Machine (SVM) algorithm and is used to build models that can predict continuous output variables. The basic principle of SVR is to find a hyperplane in a high-dimensional space that best separates the data into different classes. In the case of regression, the hyperplane is used to predict the value of the outcome variable based on the input features. Therefore, the SVR problem can be formalized as:

$$\underset{w,b}{min}\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} l_\epsilon(f(x_i), y_i) \tag{3}$$

In which, C is the regularization constant, l$\epsilon$ is the $\epsilon$-insensitive loss function. After introducing slack variables and Lagrange multipliers and taking partial derivatives, the formula of SVR can be expressed as:

$$f(x) = \sum_{i=1}^{m}\left(\widehat{\alpha}_i - \alpha_i\right)\kappa\left(x_i^T x\right) + b \tag{4}$$

**where $\kappa\left(x_i^T x\right)$=$\varnothing(x_i)^T\varnothing\left(x_j\right)$ is the kernel function.**

**Random Forest(RF)**

RF [13] is a popular ensemble learning algorithm used for classification, regression, and other machine learning tasks. The algorithm combines multiple decision trees to create a "forest" of trees that work together to make predictions. In regression problems, the output of each decision tree is averaged to obtain the final regression result [14]. The specific idea is as follows:

(1) Assuming that the training dataset contains N data objects, a training dataset is constructed by randomly sampling M samples with replacement using the bootstrap method, where each sample is not completely identical to the others.
(2) Assuming that each sample data has X features, a subset of x (x $<\ =$ X) features is randomly selected from all the features, and the best splitting attribute is chosen as the node to grow the CART decision tree, with k remaining constant during the tree growing process.
(3) Repeat the above steps to build n CART trees, and obtain the final prediction by averaging the outputs of these decision trees.

$$\widehat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b\left(x'\right) \tag{5}$$

**eXtreme Gradient Boosting(XGBoost)**

XGBoost [15] is a highly efficient gradient boosting decision tree algorithm that uses the ensemble idea - the Boosting idea - to integrate multiple weak learners into a strong learner through a certain method. Its algorithmic process is as follows:

(1) Set the model to begin with a constant value:

$$\widehat{f}_{(0)}(x) = \underset{\theta}{\arg min}\sum_{i=1}^{N} L(y_i, \theta) \tag{6}$$

(2) Calculate the gradients and hessians:

$$\widehat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\widehat{f}_{(m-1)}(x)} \tag{7}$$

$$\widehat{h}_m(x_i) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2}\right]_{f(x)=\widehat{f}_{(m-1)}(x)} \tag{8}$$

(3) Train a base learner on the training set by solving the following optimization problem:

$$\widehat{\phi}_m = \underset{\phi\in\Phi}{\mathrm{arg}min} \sum_{i=1}^{N} \frac{1}{2}\widehat{h}_m(x_i)\left[-\frac{\widehat{g}_m(x_i)}{\widehat{h}_m(x_i)} - \phi(x_i)\right]^2 \tag{9}$$

$$\widehat{f}_m(x) = \alpha\widehat{\phi}_m(x) \tag{10}$$

(4) Modify the model:

$$\widehat{f}_{(m)}(x) = \widehat{f}_{(m-1)}(x) + \widehat{f}_m(x) \tag{11}$$

(5) Output:

$$\widehat{f}(x) = \widehat{f}_{(M)}(x) = \sum_{m=0}^{M}\widehat{f}_m(x) \tag{12}$$

## 2.3 Optimization Algorithm

An optimization algorithm refers to the process of minimizing or maximizing an objective function, subject to given constraints, by finding one or more optimal or near-optimal solutions. This paper introduces a tree-structured Bayesian optimization algorithm to tune hyperparameters of the production enhancement effect prediction model. This method solves the problem of obtaining the optimal prediction model through manual tuning, providing a more efficient and effective approach.

**Tree-structured Parzen Estimator(TPE)**

TPE [16] uses two density functions to define $p(x|y)$:

$$p(x|y) = \begin{cases} l(x) & \text{if}\, y < y^* \\ g(x) & \text{if}\, y \geq y^* \end{cases} \tag{13}$$

In the above equation, $l(x)$ is established using the observation space $\{x^{(i)}\}$ and the corresponding loss $f(x^{(i)})$ is less than y*, while $g(x)$ is established using the remaining observations. The TPE-based method relies on a value of y* greater than the best observed value of $f(x)$, so that some points can be used to build $l(x)$. TPE adopts expected improvement (EI) as the acquisition function. However, since it is impossible to obtain the posterior probability $p(x|y)$, Bayesian formula is employed to transform the acquisition function:

$$\mathrm{EI}_{y^*}(x) = \int_{-\infty}^{y^*}(y^* - y)p(y|x)dy = \int_{-\infty}^{y^*}(y^* - y)\frac{p(x|y)p(y)}{p(x)}dy \tag{14}$$

In this equation, y* represents a threshold value. Let $\gamma = p(y < y^*)$ denote a certain quantile used in the TPE algorithm to partition $l(x)$ and $g(x)$. The value of $\gamma$ is in the range of (0, 1). The final simplified formula is:

$$EI_{y^*}(x) = \frac{\gamma y^* \updownarrow(x) - \updownarrow(x)\int_{-\infty}^{y^*} p(y)dy}{\gamma \updownarrow(x) + (1-\gamma)g(x)} \propto \left(\gamma + \frac{g(x)}{\updownarrow(x)}(1-\gamma)\right)^{-1} \tag{15}$$

# 3 Experiment and Result

## 3.1 Introduction to the Dataset

This study collected data on all oil production enhancement measures implemented in a certain block of an oilfield from 2017 to the present, including acidification, unclogging, and water flooding. After selecting wells where the measures were effective and conducting data cleaning and correlation analysis, the sample size of wells subjected to acidification and unclogging measures was too small to support machine learning analysis. Therefore, this study ultimately chose water flooding measures as an example for oil production enhancement prediction. The sample database contains 147 wells that achieved oil production enhancement after water flooding measures were implemented. The input variables include eight geological static features, four production dynamic features, and two measure process features, as shown in Table 1.

**Table 1.** Feature presentation table.

| Feature category | Feature name |
|---|---|
| Geological static features | Porosity of injection well, Permeability of injection well, Thickness of injection well, Oil temperature of injection well, Viscosity of crude oil of injection well, Permeability of production well, Porosity of production well, Distance between wells |
| Production dynamic features | Injection well pressure before the measure, Injection well daily volume before the measure, Daily fluid production before the measure, Daily oil production before the measure |
| measure process features | Measure chemicals, Chemicals injection volume |

According to statistics, there are three different chemical types A, B and C in 147 profile control measures samples, and the sample sizes of different classes of chemicals are significantly different, among which 96 are of type A, 33 are of type B and 18 are of type C. Unbalanced sample distribution has a great influence on the learning and prediction of machine learning model. Therefore, this study adopted the data enhancement method of oversampling to expand the data, so as to ensure the same sample size of the three measures. Generate new 410 after SMOTE oversampling and screening and use this new 410 as the data set for the forecast model.

## 3.2 Experimental Setting

To select the optimal predictive model for effects of oil-increasing measures, this study compared three commonly used machine learning algorithms for regression problems in petroleum engineering: Support Vector Regression (SVR), Random Forest (RF), and XGBoost. Predictive models were constructed for each algorithm and their performance was evaluated under multiple loss functions. The best-performing predictive model was determined, and an optimization algorithm was introduced to fine-tune the model's hyperparameters, further enhancing its predictive accuracy.

### 3.3 Result

Based on three different algorithms, data-driven predictive models were developed for oil-increasing measures using geological static parameters, production dynamic parameters, and process parameters as inputs, and post-measure oil production as output. The training and testing sets were divided in a 9:1 ratio. The prediction results of the different algorithms are shown in Fig. 1. By comparing the performance of the models using the same testing set, it can be observed that the predictive accuracy of XGBoost and RF algorithms are significantly higher than that of SVR algorithm.

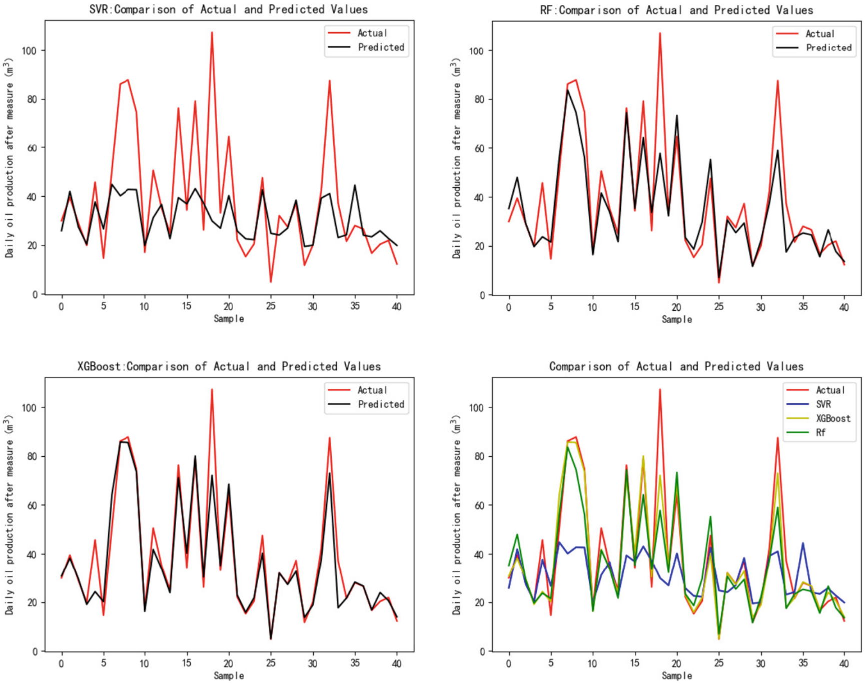


**Fig. 1.** The effect comparison of different algorithms

To conduct a precise analysis of the prediction performance of RF and XGBoost algorithms, this study comprehensively evaluated their performance using three loss functions: Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R2), as shown in Table 2. From the numerical results, it can be seen that XGBoost algorithm outperforms the other two algorithms. Therefore, XGBoost was selected for further research.

**Table 2.** Evaluation of three algorithms

| Algorithms | MAE | MSE | $R^2$ |
|---|---|---|---|
| SVR | 13.12 | 453.82 | 0.26 |
| RF | 7.09 | 136.76 | 0.78 |
| XBGoost | 4.39 | 68.22 | 0.88 |

The above model obtained the optimal results by manually adjusting the hyperparameters after determining their approximate range using grid search. However, manual tuning of hyperparameters can hardly result in the best combination of model parameters, and there is still room for improvement in hyperparameter performance. Therefore, this study introduced a Bayesian optimization algorithm based on Tree-structured Parzen Estimator (TPE) to optimize the hyperparameters of the XGBoost prediction model. The final optimization result returned the maximum value of R2, and the optimized parameters are shown in Table 3.

**Table 3.** Hyperparameter optimization results

| Hyperparameter | optimization results | Hyperparameter | optimization results |
|---|---|---|---|
| max_depth | 5 | n_estimators | 498 |
| learning_rate | 0.0581 | gamma | 6.1491 |
| colsample_bytree | 0.9185 | subsample | 0.5102 |
| reg_alpha | 0.0196 | reg_lambda | 1.5302e-07 |

The performance of the TPE-XGBoost model for predicting the effect of enhanced oil recovery measures, incorporating the optimization algorithm, is shown in Fig. 2 and Table 4. It can be observed that the introduction of the optimization algorithm improves the performance of the predictive model, with the optimized model outperforming the non-optimized model under all three loss functions. The final predictive accuracy ($R^2$) can exceed 90%.

**Table 4.** The effect after hyperparameter optimization

| Algorithms | MAE | MSE | $R^2$ |
|---|---|---|---|
| XBGoost | 4.39 | 68.22 | 0.88 |
| TPE-XGBoost | 4.11 | 58.19 | 0.91 |

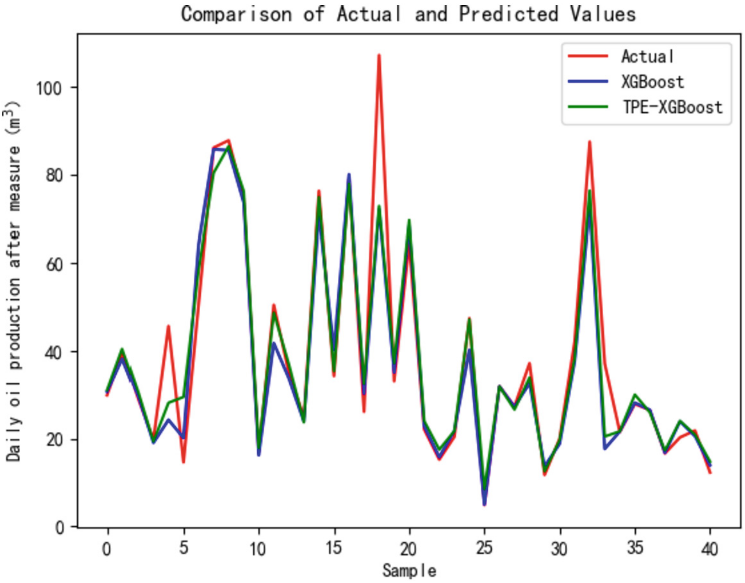


**Fig. 2.** The effect after hyperparameter optimization

## 4 Conclusion

This article proposes a data-driven method for predicting the effects of oil-increasing measures based on the TPE-XGBoost algorithm. This method first enhances the data samples to some extent, which alleviates the problem of insufficient sample size. At the same time, the model comprehensively considers three types of features: geological static parameters, production dynamic parameters, and measure process parameters, deeply mining their relationship with the effect of increasing oil, and automatically optimizing the model hyperparameters to achieve the prediction of daily oil production after the measures, which significantly improves the prediction accuracy compared with other algorithms and can reach over 90%. However, the current research is still limited by the insufficient quality of on-site data. In future research, in addition to obtaining high-quality data from the source, high-level feature engineering will also be the next research focus. In addition, incorporating economic indicators into machine learning is also a future direction of research.

**Acknowledgments.** This work is supported by the National Natural Science Foundation of China under Grant 52274057, 52074340 and 51874335, the Major Scientific and Technological Projects of CNPC under Grant ZD2019–183-008, the Major Scientific and Technological Projects of CNOOC under Grant CCL2022RCPS0397RSN, the Science and Technology Support Plan for Youth Innovation of University in Shandong Province under Grant 2019KJH002, 111 Project under Grant B08028.

## References

1. Chen, S.Y., Zhang, Q., Mclellan, B., et al.: Review on the petroleum market in China: history, challenges and prospects. Pet. Sci. **17**, 1779–1794 (2020)
2. Gong, X., Sun, Y., Du, Z.: Geopolitical risk and China's oil security. Energy Policy **163**, 112856 (2022)
3. Alfarge, D., Wei, M., Bai, B.: Evaluating the performance of hydraulic-fractures in unconventional reservoirs using production data: comprehensive review. J. Nat. Gas Sci. Eng. **61**, 133–141 (2019)
4. Zhang, Q., Zhu, W., Liu, W., Yue, M., Song, H.: Numerical simulation of fractured vertical well in low-permeable oil reservoir with proppant distribution in hydraulic fracture. J. Petrol. Sci. Eng. **195**, 107587 (2020)
5. Hassan, A.M., Aljawad, M.S., Mahmoud, M.A.: Predicting the productivity enhancement after applying acid fracturing treatments in naturally fractured reservoirs utilizing artificial neural network. In: Abu Dhabi International Petroleum Exhibition & Conference. OnePetro (2021)
6. Vaishya, R., Javaid, M., Khan, I.H., Haleem, A.: Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes Metab. Syndr. **14**(4), 337–339 (2020)
7. Sircar, A., Yadav, K., Rayavarapu, K., Bist, N., Oza, H.: Application of machine learning and artificial intelligence in oil and gas industry. Petrol. Res. **6**(4), 379–391 (2021)
8. Koroteev, D., Tekic, Z.: Artificial intelligence in oil and gas upstream: trends, challenges, and scenarios for the future. Energy AI **3**, 100041 (2021)
9. Xue, L., Liu, Y., Xiong, Y., Liu, Y., Cui, X., Lei, G.: A data-driven shale gas production forecasting method based on the multi-objective random forest regression. J. Petrol. Sci. Eng. **196**, 107801 (2021)
10. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. Big Data **6**(1), 1–48 (2019)
11. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
12. Awad, M., Khanna, R.: Support vector regression. Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers, 67–80 (2015)
13. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
14. Grömping, U.: Variable importance assessment in regression: linear regression versus random forest. Am. Stat. **63**(4), 308–319 (2009)
15. Chen, T., Guestrin, C.: Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
16. Bergstra, J., Bardenet, R., Bengio, Y., Kégl, B.: Algorithms for hyper-parameter optimization. In: Advances in Neural Information Processing Systems, vol. 24 (2011)

# Research on Stability Evaluation and Adjustment Method of Concentrated Casing Damage Area Based on Gradient Lifting Algorithm

Bing-bing Yang[1](✉), Ji-yuan Lu[1], You-chun Wang[1], Peng-chao Xu[2], Shu-juan Zhang[1], Xue-yan Jiang[1], and Li-qiu Zhang[1]

[1] Exploration and Development Research Institute of Daqing Oilfield Co. Ltd., Daqing, Heilongjiang, China
524911612@qq.com

[2] No. 1 Oil Production Plant of Daqing Oilfield Co. Ltd., Daqing, Heilongjiang, China

**Abstract.** Casing damage is a common problem faced in the development of old oilfields. Some areas with weak mechanical properties of rock layers and other unfavorable factors are prone to severe casing damage, such as time of occurrence and plane distribution of casing damage wells. The stable state of damaged rock formations is a prerequisite for the implementation of workover and remaining oil tapping. Traditional stability evaluation of damaged areas often relies on qualitative analysis through well condition investigation and engineering logging, with incomplete considerations and a high rate of judgment errors, often leading to secondary concentrated casing damage after treatment. Therefore, based on the research and understanding of centralized casing damage mechanism, taking into account the changes in casing damage situation and the control of inducing factors, 9 indicators are selected to establish an evaluation index system for stability of casing damage areas. Gradient lifting algorithm is applied to achieve quantitative grading evaluation of the stability, with a verification compliance rate of 83.3%; meanwhile, classified adjustment measures are implemented to overcome unstable aspects, shorten the adjustment cycle, determine the timing of overall governance as soon as possible, and ensure the oil recovery of casing damage areas. This

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 31–42, 2024.
https://doi.org/10.1007/978-981-97-0272-5_3

method was applied in the X67 casing damage area, guiding the implementation of various stability control workloads for 49 wells, and effectively improving the stability of the casing damage area.

**Keywords:** concentrated casing damage area · comprehensive assessment · stability adjustment · gradient lifting algorithm

## 1 Introduction

The casing damage in old sandstone oilfields developed by water injection is severe. There is a phenomenon of concentrated casing damage, in some areas with large dip angles, complex fractures, and large pressure differences [1–5]. The time of casing damage occurrence, the location of casing damage layers, and the plane distribution of casing damage wells are concentrated, which poses great harm to oilfield development effectiveness, cost reduction and efficiency improvement, as well as safety and environmental protection. Due to a lack of scientific understanding of the stability of casing damage areas, production sites often rely on operational surveys or engineering monitoring to qualitatively analyze the stability of casing damage areas based on information such as the number of newly discovered casing damage wells and changes in inner diameter of damaged casing [6]. This has led to improper timing of governance determined in some concentrated casing damage areas, resulting in a recurrence of concentrated casing damage after governance, seriously affecting the oil recovery process of these blocks and significantly increasing the cost of well repair operations [7–10]. At present, there are still two concentrated casing damage areas in Daqing Changyuan Oilfield that have not fully resumed production and urgently need further treatment. In order to scientifically determine the timing of casing damage area treatment and restore the production of casing damage area as soon as possible, there is an urgent need for an accurate and quantitative centralized casing damage area stability evaluation and adjustment method, which comprehensively evaluates the stability status of the damaged layer in the casing damage area, and implement adjustments against unstable factors and causes to promote the stability of the casing damage area as soon as possible, ensure the comprehensive treatment effect of the casing damage area, and avoid the occurrence of secondary concentrated casing damage situations.

This article is based on the research and understanding of the mechanism of concentrated casing damage. Mainly from two aspects, the trend changes in casing damage situation and the inducing factors that affect the stability of damaged formation layers, a stability evaluation index system and thresholds for concentrated casing damage areas are created. Gradient lifting algorithms are optimized, and comprehensive evaluation of casing damage area stability is carried out. At the same time, adjustment measures against unstable factors in casing damage areas are provided to improve the stability of damaged rock layers, providing guidance for the reasonable timing determination of well workovers and drillings.

## 2 Mechanism of Concentrated Casing Damage

In Daqing Changyuan Oilfield, there were 23 concentrated casing loss zones during the three casing loss peak periods, of which 22 were concentrated casing loss zones of maker bed as N_2 bottom. Through combing and comparing the causes of historical casing loss in Table 1, the following characteristics were found in most of the concentrated casing loss zones.

**Table 1.** Cause classification of concentrated casing damage zone

| Geological Factors | Development Factors | Engineering Factors |
|---|---|---|
| complex structure<br>complicated fault<br>large dip angle<br>dense fossils | improper process before and after drilling<br>uneven injection and production<br>large pressure differences<br>large pressure rise during flooding | poor cementing quality<br>renew with incomplete plugging<br>untimely casing leap discovery<br>abnormal water absorption at the top of perforation |

Firstly, in terms of geological factors, wells with casing damage are mostly located in the area with large dip angle, the complex faults and the layers rich in fossil, especially the oil shale in the maker bed of the N_2 bottom with well-developed horizontal joints and micro-fractures, which are easy to crack. The Fig. 1 shows that micro-fractures expand after water immersion, and the Fig. 2 shows that the rock strength decreases by 16.4%, which further weakens the shear resistance.

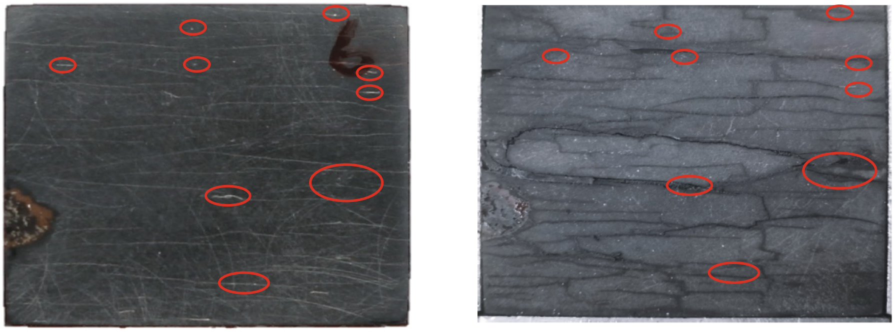
Secondly, in the development process, under the influence of unbalanced injection and production adjustment or unreasonable drilling off and restoring, large regional pressure difference or pressure change is generated, which leads to the sliding shear along the bedding surface of the N_2 bottom oil shale, finally resulting in a large area of concentrated casing losses.

At last, the engineering factors, due to some cementing with poor quality, untimely casing leap discovery and incomplete well plugging, injected water get into the N_2 bottom along the channeling space, when there is abnormal water absorption at the top section of the perforating, resulting in rock mechanical properties deteriorate, and ultimately concentrated casing damage.

It can be seen from the above that the concentrated casing failure of the N_2 bottom is caused by its mechanical properties, the influence of water immersion and the external abnormal pressure. Therefore, the stability of the N_2 bottom maker bed is not only affected by its own lithology and mechanical properties, but also affected by abnormal development and water immersion factors.

## 3 Evaluation Method for Stability of Casing Damage Zone

Based on the understanding of the mechanism of concentrated casing loss, in addition to the analysis of the current casing loss situation, the scientific and comprehensive evaluation for the stability of the formation layer with concentrated casing loss, should



**a.** Original sample

**b.** Sample after water immersion

**Fig. 1.** Fracture extension induced by water immersion in the black shale

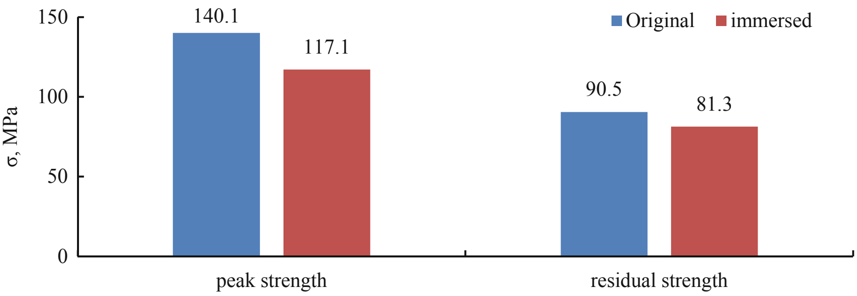


**Fig. 2.** Changes in rock strength of N_2 bottom layer before and after immersion

focus on the development anomaly and the risk factors of water immersion that will affect the stability of the rock layer in the future. Therefore, the stable state of the casing loss area should be divided into two aspects. First, the casing loss situation is no longer aggravated, which is mainly manifested as the area of the casing loss area does not expand, the number of casing loss wells does not increase, and the casing loss degree does not worsen. Secondly, the inducement factors of the centralized casing loss are controlled or eliminated, including the water immersion risk of the N_2 bottom and the abnormal development risk, to ensure that the secondary centralized casing loss will not occur after the recovery of treatment.

### 3.1 Stability Evaluation Index System

Starting from the two aspects of stability evaluation in the cased damage area, 23 evaluation indexes are selected preliminary, including 8 indexes related to the cased damage situation, 10 for water immersion risk, and 5 for abnormal development risk.

Spearman correlation coefficient method was used to quantify the relationship between 23 stability evaluation indexes and the occurrence of centralized casing loss.

Among them, if the first correlation coefficient is greater than 0.7, it is strong correlation; if the correlation coefficient is less than 0.4, it is weak correlation; and the rest are medium correlation. Indicators with strong correlation are taken as alternative indicators.

**Table 2.** Preliminary design of stability evaluation index of casing damaged area

| Classification | Trend change index | Inducing factors | |
|---|---|---|---|
| | | Risk of water immersion | Abnormal development indexes |
| detailed index | change rate of casing damage area<br>rate of casing damage found during jobs<br>annual casing damage rate<br>annual casing loss rate of N_2 bottom layer<br>ratio of leap wells<br>change rate of inner diameter of casing<br>change in stress direction of casing damage well<br>time interval between discovery of casing damage and last job | abnormal drilling information<br>updated numbers of incompleted abandon injection well<br>proportion of poorly cementing wells<br>proportion of water immersion risk wells<br>water immersion area<br>proportion of producing liquid<br>proportion of abnormal testing<br>ratio of unloading wells to water immersion wells<br>ratio of abnormal injection wells<br>centralized investigation | risk level of block in casing damage<br>proportion of wells with casing damage risk<br>pressure increasing rate in early stage of chemical flooding<br>inject-pro pressure difference in the middle and later stages of chemical flooding<br>injection rate of chemical flooding |
| quantity | 8 | 10 | 5 |

Spearman correlation coefficient method was adopted to calculate the correlation coefficient among alternative indicators. If the correlation coefficient between the two alternative indicators is greater than 0.4, the correlation coefficient between the two alternative indicators and the casing loss rate should be referred to, and the one with the larger first correlation coefficient should be selected as the stability evaluation index. If the correlation coefficient between the two alternative indicators is less than 0.4, the two alternative indicators are retained as stability evaluation indicators.

According to the above screening methods, eight evaluation indexes that are strongly correlated with the occurrence of centralized casing loss and relatively independent of each other are obtained, which constitute the evaluation index system of the stability of the casing loss area (see Table 3).

### 3.2 Threshold of Stability Evaluation Index

An example is taken to illustrate the determination process of indicator thresholds by the ratio of water unloading wells to immersion wells. Firstly, the scatter relationship

**Table 3.** Optimized results of the evaluation index

| Classification | | Index | Unit | Definition |
|---|---|---|---|---|
| trend change | area variation | rate change of casing damage area | % | the ratio of the area difference between the current and the previous year to the damaged area zone in the previous year |
| | wells variation | rate of casing damage found during jobs | % | the proportion of newly discovered casing damage wells to the number of jobs |
| | | annual casing loss rate of N_2 bottom layer | % | the proportion of newly discovered damage wells of N_2 bottom layer to the total number of the year |
| | Degree variation | rate change of inner diameter of casing | % | the proportion of changes in the inner diameter of casing damaged compared to the original average inner diameter |
| inducing factors | risk of water immersion | proportion of water immersion risk wells | % | the proportion of wells that are scrapped and not completely renewed and have a cementing quality less than 60% in the total number of water injection wells |
| | | ratio of water unloading wells to immersion wells | 1 | the ratio of the number of reperforated producers and those with casing leap to the number of injectors with casing leap |
| | abnormal development indexes | risk level of block in casing damage | / | risk level of casing loss in block |
| | | proportion of wells with casing damage risk | % | the proportion of injectors with casing loss risk in the number of injectors |

between index data and the casing loss rate was established. The casing loss rate of 3% was taken as the threshold, and the horizontal coordinate was divided into several intervals with 0.4 step length (see Fig. 3). The proportion of blocks with casing loss rate greater than 3% in different intervals was calculated, and the value corresponding to the obvious inflection point on the curve was selected as the threshold of the index (see Fig. 4). According to the above practices, the limits of other 7 evaluation indicators are determined (see Table 4), which are mainly divided into two types: the first are positive indicators, which the bigger the better; the second are reverse indicators, which the smaller the better.

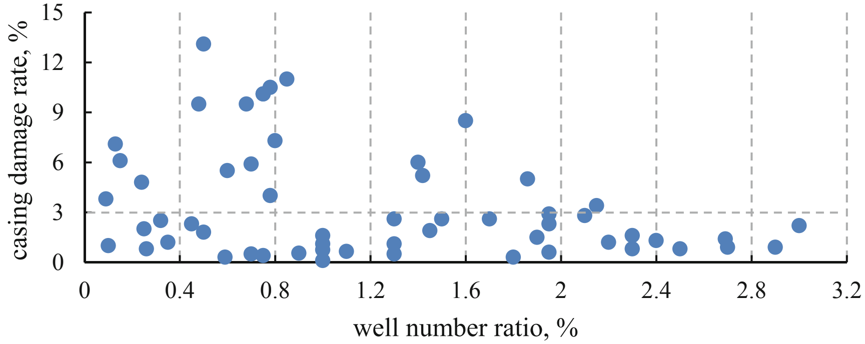


**Fig. 3.** Scatter relationship between the ratio of water unloading to immersion wells

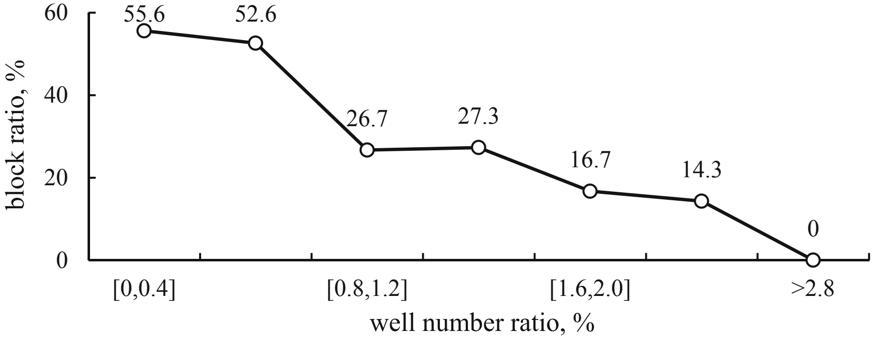


**Fig. 4.** The proportion of blocks with casing damage rate > 3% in each well ratio interval

### 3.3 A Set of Stability Evaluation Methods

**Grading Standards of Stabilities.** According to the definition of casing loss zone stability evaluation, the stability of casing loss zone is divided into three levels: stable, understable and unstable.

**Table 4.** Thresholds and controlling range of the index for stability evaluation

| Classification | | Index | Unit | Control range |
|---|---|---|---|---|
| trend change | change of area | change rate of casing damage area | % | [0, 1.0] |
| | change of wells | rate of casing damage found during job | % | [0, 3.5] |
| | | annual casing loss rate of N_2 bottom layer | % | [0, 1.0] |
| | change of degree | change rate of inner diameter of casing | % | [0, 9.0] |
| inducing factors | risk of water immersion | proportion of water immersion risk wells | % | [0, 5.0] |
| | | ratio of unloading wells to water immersion wells | 1 | (0.8, ∞] |
| | abnormal development indexes | risk level of block in casing damage | / | medium to low |
| | | proportion of wells with casing damage risk | % | [0, 9.0] |

**Stability Classification of Historical Damage Area.** According to the stability evaluation index data of the area after the occurrence of centralized casing loss, the stability grade of the casing loss area over the years is divided according to the stability grade standard of the casing loss area in Table 5.

Spearman correlation analysis algorithm is used to calculate the correlation among different levels. The weaker the correlation is, the clearer the classification is. If the correlation between the two levels is strong (correlation coefficient greater than 0.4), the block stability is regraded, until the correlation between grades is weak or no correlation (correlation coefficient less than 0.4), in Fig. 5.

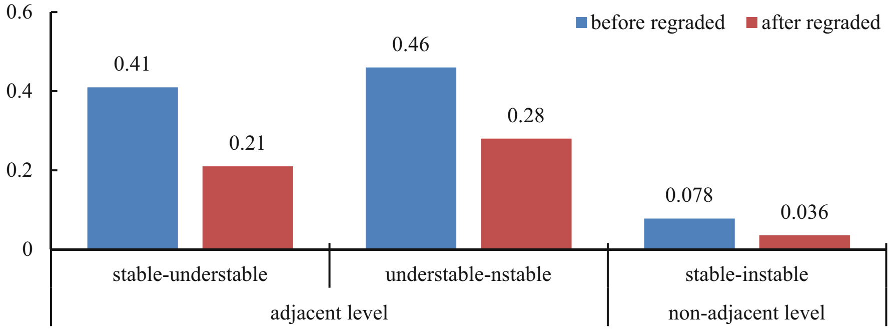


**Fig. 5.** Correlation calculation results before and after adjustment

**Optimization of Stability Evaluation Methods.** 75% of the casing loss area data was taken as sample data and 25% as verification data. A variety of big data analysis methods were used to carry out deep learning on the sample data respectively, forming the evaluation model of each method, calculating the respective and the verification coincidence rate, and selecting the method model with the model coincidence rate of over 90% and the highest verification coincidence rate of over 80% as the optimal evaluation model. In the end, the coincidence rate of 7 algorithms reached more than 90%. After the comparison and verification of the coincidence rate, the gradient lifting algorithm was selected as the evaluation method for the stability of the casing loss area (see Table 5).

**Table 5.** Comparison of coincidence rates of various evaluation methods

| Evalution Methods | Model Compliance Rate (%) | | | | Verification Compliance Rate (%) | | | |
|---|---|---|---|---|---|---|---|---|
| | Stable | Understable | Instable | Overall | Stable | Understable | Instable | Overall |
| stochastic gradient descent | 79.6 | 75.0 | 89.7 | 81.5 | 84.0 | 62.5 | 92.3 | 79.6 |
| ridge regression | 68.4 | 76.0 | 83.3 | 75 | 72.4 | 50.0 | 84.6 | 70.4 |
| logistic regression | 100 | 90.5 | 92.5 | 94.4 | 90.9 | 64.7 | 86.7 | 81.5 |
| decision tree | 100 | 100 | 100 | 100 | 90.9 | 54.6 | 90.0 | 75.9 |
| extra trees | 100 | 100 | 100 | 100 | 95.2 | 54.6 | 81.8 | 75.9 |
| random forest | 91.3 | 91.7 | 90.5 | 91.1 | 95.5 | 64.7 | 80.0 | 81.5 |
| gradient boosting | 95.5 | 92.7 | 97.4 | 95.2 | 90.9 | 68.4 | 92.3 | 83.3 |
| adaboost classification | 81.3 | 67.4 | 96.7 | 79.8 | 90.9 | 59.1 | 90.0 | 77.8 |
| support vector machine | 100 | 100 | 100 | 100 | 100 | 30.0 | 0.0 | 35.2 |
| naive_bayes | 71.8 | 41.7 | 60.7 | 60.5 | 85.7 | 35.7 | 53.9 | 57.4 |
| neural networks | 95.4 | 92.9 | 100 | 96.0 | 95.5 | 65.0 | 91.7 | 83.3 |

## 4 Adjustment Measures for Stability

According to the principle of one countermeasure for one index, 12 personalized adjustment measures are formulated according to the adjustable factors affecting the stability of the damage zone, two risk factors of water immersion into N_2 bottom layer and abnormal development, combined with the common injection and production adjustment measures on the production site. In terms of immersion risk, the main way is to reduce the water immersion area and prevent the water immersion area from further expanding by the combination method of plugging and dredging. In the aspect of abnormal development risk, the abnormal single well with casing loss risk can be restored to the limit of casing loss warning by adjusting the formation plane and the pressure difference between layers and treating them, according to Table 6.

**Table 6.** Stability adjustment measures of casing damage area

| Classification | Adjustment Measure | |
|---|---|---|
| risk of water immersion | water immersion control | cementing quality reevaluation and channeling plugging, injection interval adjustment |
| | | plugging of incompletely abandoned well or shut down the replacement well |
| | | packer seal inspection |
| | discharge storage | perforating of the production well |
| | | maintaining production with casing leap |
| abnormal development | pressure differential adjustment | adjustment of injection and production parameters for regions, well patterns, and well groups |
| | abnormal well treatment | adjust production indexes to a reasonable range, conduct inspection, leak detection |

## 5 Method Application

X67 damage zone has an oil-bearing area of 20.6 km$^2$ and geological reserves of 43.4 million tons. Concentrated damage occurred in the N_2 bottom of the tender in 2015. By the end of July 2021, 137 wells and 18.7% of wells were lost. The well opening rate was only 57.8%, the daily water injection accumulation decreased by 69.9%, and the injection-production ratio was only 0.50, with an annual decline rate of 0.97%.

The above method was used to evaluate the stability, and the results showed that the stability was unstable. In the aspect of sheath damage, it was found that the sheath loss rate of the N_2 bottom was higher. In terms of inducing factors of casing loss, the main factors are the hidden danger of water channeling caused by a large number of incomplete abandoned wells, and the high risk of secondary casing loss caused by low injection-production ratio, continuous reduction of formation pressure and large pressure difference.

In order to improve the stability of the casing damage zone, 49 adjustments were implemented according to the principle of overall planning, coordination and classification, aiming at the risk of water immersion into the N_2 bottom layer and the unstable factors that could easily exacerbate the casing damage in terms of development, from the perspective of improving the overall stability of the casing damage zone. Since it was implemented for more than half a year, the risk of water immersion has been effectively controlled, the difference of plane and inter-layer pressure has been reduced, no testing and operation anomalies have occurred during this period, the casing damage situation has been significantly improved, and the stability of the casing damage area has been improved to a certain extent.

## 6 Conclusion

From the two aspects, the situation of the damage zone and the control of inducing factors, the stability evaluation index is designed and optimized to quantitatively characterize the current situation of the damage zone and the potential risks in the future, which can realize the scientific and comprehensive evaluation of the stability of the damage zone, and provide a method to scientifically guide the comprehensive treatment of the damage zone, determine the reasonable treatment time, and avoid the secondary concentrated damage after treatment.

In view of the unstable factors of the casing loss area, timely adjustment measures can effectively improve the stability of the casing loss area, shorten the treatment period, and restore the production contribution as soon as possible.

## References

1. Liu, S., Zhang, H.: Study on the threshold of injection pressure in dense fault-dense zone. Dril. Prod. Technol. **43**(04), 51–53+8–9 (2020)
2. Wang, D.: Reasons and analysis of casing damage in Zhongyuan oilfield. Pet. Drill. Tech. (02), 36–38 (2003)
3. Ren, L., Wang, F., Dou, S., et al.: Preliminary understanding of casing loss law in Dagang Zaoyuan oilfield. Drill. Prod. Technol. (03), 127–128+133 (2008)
4. Zhang, J.: Study on causes and prediction methods of casing damage in complex faulted basins. Ocean University of China (2014)
5. Ye, C., Zhan-Lin, Li, P.: Research on casing damage law and protective measures in Karamay oilfield. Pet. Eng. Constr. (05), 11–13+4–5 (2008)
6. Liu, H., Sun, F., Chen, Q., et al.: Petroleum geology & oilfield development in Daqing (01), 26–29+5 (1993)
7. Liu, H., Liu, J., Zhuo, S., Jin, Y.: Geological factors for controlling damage of batch casing in the member II of Nenjiang Formation in Daqing Oilfield. Acta Petrolei Sinica (05), 135–138+142 (2006)

8. Zhang, S.-J., et al.: Study on causes and preventive measures of concentrated casing damage in XI area. In: Lin, J. (ed.) IFEDC 2019. SSGG, pp. 2768–2775. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-2485-1_254
9. Lu, J.-Y., Zhu, L.-H., Zhang, S.-J., Yang, B.-B., Wang, Y.-C.: Study on stability evaluation method of concentrated casing damage areas in the N_2 bottom layer in Lasa Xing oilfield. In: Lin, J. (ed.) IFEDC 2019. SSGG, pp. 2835–2844. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-2485-1_260
10. Sun, Z., Chen, T., Zhu, L., et al.: Analysis of the upper and lower boundaries of permeability evolution during shale rock shear deformation. Energy Fuels **36**(4), 2007–2022 (2022)

# A Knowledge Base of Shale Gas Play and Its Application on EUR Prediction by Integrating Knowledge Graph and Automated Machine Learning Techniques

Xiang-guang Zhou[1](✉), Rong-ze Yu[2], Wen-kuang Wu[1], and Wei Xiong[2]

[1] Information Center, Research Institute of PetroChina Exploration and Production,
Beijing, China
zhouxg69@petrochina.com.cn

[2] Unconventional Research Center, Research Institute of PetroChina Exploration
and Production, Beijing, China

**Abstract.** The objective of this study is to analyze dominant controlling factors of the EUR of shale gas wells and then to forecast the EUR precisely by employing knowledge graph and automated machine learning techniques. First, an ontology knowledge representation model and a set of classification system for shale gas production are constructed, which include 13 shale gas objects such as basin, shale gas play, shale gas field, shale gas reservoir, and shale gas well, and their 112 geological, engineering and production parameters, such as mineral brittleness, fracturing section length, sanding intensity, and first-year production, and so on. Subsequently, structured data from existing databases are transformed, and loaded into the knowledge base. Large amount of unstructured data from papers, presentations, professional books are extracted and loaded by using various natural language processing (NLP) tools. The final shale gas knowledge base contains 56 shale plays and more than 1,000 shale gas wells worldwide. Based on the shale gas knowledge base, the graph embedding algorithm is used to convert the graph into a vector in order to train the machine learning models. Various automated machine learning frameworks such as TPOT, H2O, Auto-Sklearn, and AutoGluon are implemented and the performances are compared. According to

23, IFEDC Organizing Committee.

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20–22 September 2023.

This paper was selected for presentation by the IFEDC Committee following review of information contained in an abstract submitted by the author(s). Contents of the paper, as presented, have not been reviewed by the IFEDC Technical Team and are subject to correction by the author(s). The material does not necessarily reflect any position of the IFEDC Technical Committee its members. Papers presented at the Conference are subject to publication review by Professional Team of IFEDC Technical Committee. Electronic reproduction, distribution, or storage of any part of this paper for commercial purposes without the written consent of IFEDC Organizing Committee is prohibited. Permission to reproduce in print is restricted to an abstract of not more than 300 words; illustrations may not be copied. The abstract must contain conspicuous acknowledgment of IFEDC. Contact email: paper@ifedc.org.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 43–54, 2024.
https://doi.org/10.1007/978-981-97-0272-5_4

the model with best performance, the main controlling factors of the EUR of shale gas wells are high-quality bed thickness, fracturing section length, and fracturing fluid volume, etc., which are consistent with shale gas production practices. The MSE and MAE of the best model on the testing dataset are 0.06 and 0.19, respectively. The approach of knowledge base construction and application developed in this paper can be extended to the entire life cycle of E&P process, which can make full use of various documents, data and knowledge accumulated in the oil and gas industry to conduct decision support.

**Keywords:** Knowledge Graph · Shale Gas Production · EUR Prediction · Automated Machine Learning

## 1 Introduction

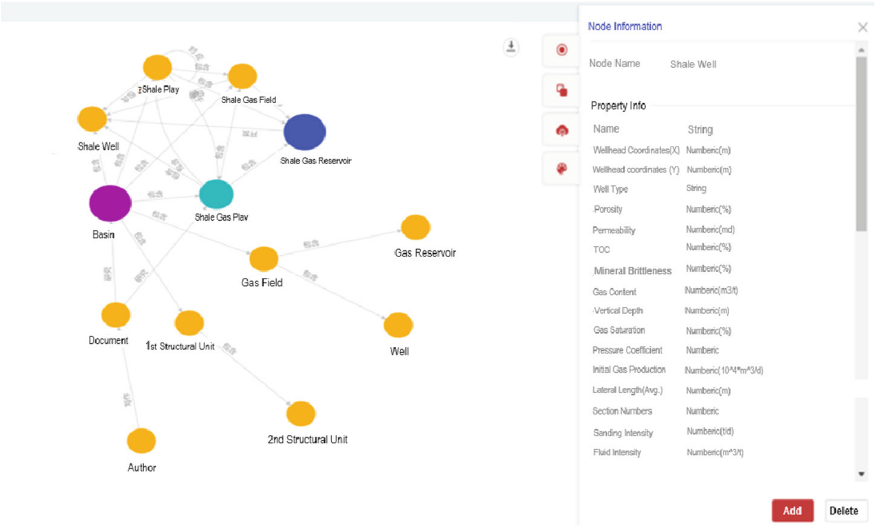
Shale gas has become an increasingly important source of energy, and its production has seen significant growth over the past decade. However, the production of shale gas wells can be challenging to forecast due to the complex geological structures, engineering, and production characteristics of shale formations. In recent years, many studies on shale gas production forecasting have been carried out and there have been developed various methods to forecast the production of the shale gas, which can be divided into three categories, empirical formulas, analytical models, and numerical formulas [12]. Each method has its advantage and can be applicable for a few shale gas wells in specific areas. Therefore, new methods are needed in order to predict the production of shale gas effectively and efficiently with high performance.

Knowledge graph and machine learning techniques have emerged as powerful tools for collecting more data from both structured and unstructured data sources, analyzing large amounts of data, and making accurate predictions. In this paper, we will explore how these techniques can be used to forecast the production of shale gas wells. Knowledge Graphs (KGs) have emerged as a compelling abstraction for organizing the world's structured knowledge, and to integrate information extracted from multiple data sources. Knowledge graphs have started to play a central role in representing the information extracted using natural language processing and computer vision. Domain knowledge expressed in KGs is being input into machine learning models to produce better predictions [3].

## 2 Shale Gas Knowledge Base Construction

### 2.1 Ontology Construction of Shale Gas Knowledge Base

First of all, it is necessary to construct a comprehensive shale gas knowledge base to store production parameters and data such as shale gas geology, structure, drilling and acid fracturing. The process can be divided into two steps. The first step is to create shale gas ontologies, and the second step is to fill in shale gas entities according to the structure of ontologies using various data sources. Ontology is an abstraction or concept of things in the real world. It usually refers to a collection of entities with the same characteristics. To be specific, business experts abstract the process of shale gas exploration, development, and production to identify shale gas ontologies such as basins, shale plays, shale gas fields, shale gas reservoirs, shale gas wells, wellbores, and fracturing stages. These objects contain many attributes or features. For example, a basin can be described by attributes such as basin name, structure type, country to which the basin belongs, maximum heat flow, minimum heat flow, oldest age of the basin, sedimentary thickness, age of the stratum, and oil and gas prospects of the basin, while the attributes of shale gas play include zone area, zone location, average reservoir burial depth, zone type, reserve abundance, recovery factor, depositional environment, brittle mineral content, pressure coefficient, reservoir thickness, permeability, porosity, ratio of adsorbed gas, gas content, TOC, gas saturation, etc. At the same time, there are various relationships between each ontology in the shale gas knowledge base. For example, a basin may contain shale gas plays, a shale gas well is located in a shale gas play, and a shale gas well includes latitude, longitude, well type, fracturing formation, surface elevation, bushing elevation, spud date, completion depth, completion date, completion formation, completion type, depth of maximum well deviation, fracturing section length, effective porosity, mineral brittleness, pressure coefficient, formation temperature, lateral length, cluster spacing, fracturing fluid strength, sanding strength, TOC and other properties (see Fig. 1).



Node Name Shale Well

Property Info

| | |
|---|---|
| Name | String |
| Wellhead Coordinates(X) | Numberic(m) |
| Wellhead coordinates (Y) | Numberic(m) |
| Well Type | String |
| Porosity | Numberic(%) |
| Permeability | Numberic(md) |
| TOC | Numberic(%) |
| Mineral Brittleness | Numberic(%) |
| Gas Content | Numberic(m3/t) |
| Vertical Depth | Numberic(m) |
| Gas Saturation | Numberic(%) |
| Pressure Coefficient | Numberic |
| Initial Gas Production | Numberic(10^4*m^3/d) |
| Lateral Length(Avg.) | Numberic(m) |
| Section Numbers | Numberic |
| Sanding Intensity | Numberic(t/d) |
| Fluid Intensity | Numberic(m^3/t) |

**Fig. 1.** A schematic diagram of shale gas ontologies, illustrating the attributes and relationships of the ontologies.

Using knowledge graph technology to model shale gas development and production process, compared with traditional relational database, the biggest advantage is the flexibility and usability, which means shale gas professionals do not need to consider all of entities, attributes, and the relationship among them in advance. The data model of the knowledge graph can be maintained or updated at any time according to actual business requirements, and the application based on the database has little impact. According to the actual business situation, the study sorts out and creates 13 types of ontologies with 112 attributes (see Fig. 1).

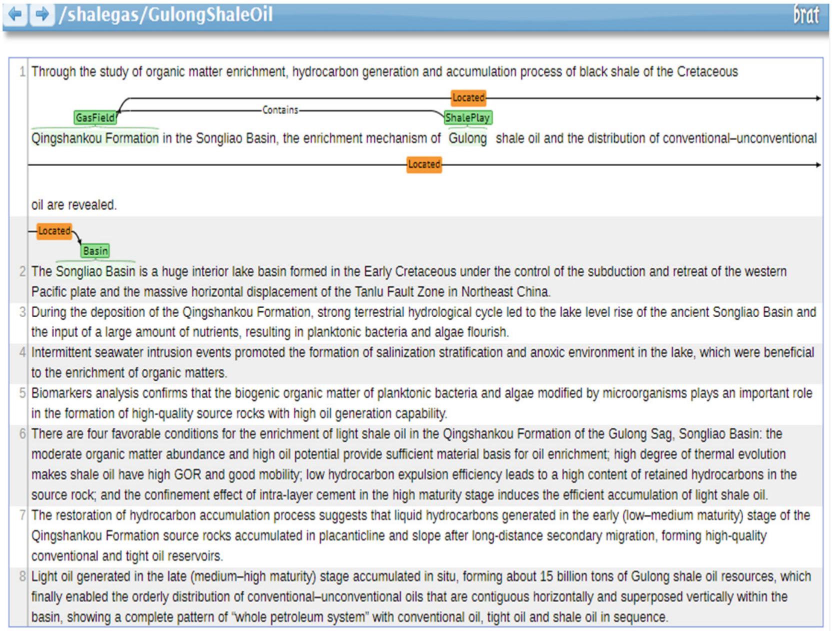
After the shale gas ontologies have been constructed, the shale gas entity can be filled in. Entity refers to something that is distinguishable and exists independently in the world. For example, Permian Basin, Eagle Ford Shale Play, MIP 3H Well, and so on are entities in the field of shale gas exploration and production. Everything in the world is composed of specific objects, that is, entities. Entities are the most basic elements in knowledge graphs, and different entities have different relationships. Every entity belongs to a specific ontology. For example, Permian Basin is a basin, which means the entity of the Permian Basin belongs to the ontology of basin. This entity has various attributes defined by the ontology, such as basin name, structure type, country to which the basin belongs, maximum heat flow, basin, minimum heat flow, the oldest age of the basin, sedimentary thickness, etc. Various shale gas entities, attributes, and relationships between entities constitute the shale gas knowledge base.

## 2.2 Entity Extraction from Shale Gas Knowledge Base

Once the ontology of the shale gas knowledge base is constructed, it is necessary to collect various types of data to fill in the knowledge base in order to expand the scale of the knowledge base. Entity extraction includes steps such as entity extraction, relationship extraction, and attribute extraction. This study collected a large amount of data, which can be divided into structured data, semi-structured data, and unstructured data in terms of data source types. Structured data refers to data that can be logically expressed in two-dimensional tables, such as Excel data tables, tables in databases, etc.

Semi-structured data is a form of structured data that does not conform to the data model structure associated with relational databases or other data tables, but contains relevant tags to separate semantic elements and to classify records and fields. Layer, the structure and content of the data are mixed, there is no obvious distinction, therefore, it is also called a self-describing structure, simply speaking, semi-structured data is between fully structured data and completely unstructured data. For example, HTML documents, JSON, XML and some NoSQL databases are semi-structured data. Unstructured data is data without a fixed structure, including various formats of office documents, text, pictures, various reports, images, and audio/video information.

For structured data, the column name of the structured data is mapped to the attribute name of the graph database entity through Mapping, and the value of the column name is filled into the attribute value of the knowledge base entity. For unstructured data, we used a rule-based, natural language model combined with manual review to extract shale gas entities, relationships, and attribute values. The task of entity extraction is to find named entities from unstructured documents. In this study, entities such as basins, shale gas zones, shale gas reservoirs, shale gas wells, wellbores, and fracturing sections are mainly extracted. For example, the system uses the preset regular expression "[AZ].{2,20}Shale[]{0,1}[Gas]{0,1}[]{0,1}[Play]{0,1}" will identify and extract "Eagle Ford Shale Play", "Bakken Shale Play", "Barnett Shale", "Fayetteville Shale", "Haynesville Shale", "Marcellus Shale", "SiChuan Shale" and other shale gas plays; for the relationship between entities, attributes and attribute values, this paper extracts knowledge from unstructured documents through manual annotation. Shale gas professionals marked the 7 types of shale gas entities, relationships, attributes and attribute values involved in this research through the annotation tools (see Fig. 2). The marked results can be directly imported into the shale gas knowledge base to form shale gas knowledge triples. Meanwhile, it provides training corpus for the natural language processing extraction model. Based on the annotation results, this study trains the Bert-BiLSTM-CRF natural language processing extraction model [5, 9], and fine-tunes the parameters according to the characteristics of the shale gas corpus. After the training is completed, the accuracy rate of shale gas entity extraction on the testing dataset reaches 86%, and the recall rate is 79%. The trained model can be applied to new shale gas documents to extract entities, relationships, attributes, and attribute values automatically or semi-automatically, which can be loaded into the shale gas knowledge base after manual reviewing.



1 Through the study of organic matter enrichment, hydrocarbon generation and accumulation process of black shale of the Cretaceous Qingshankou Formation in the Songliao Basin, the enrichment mechanism of Gulong shale oil and the distribution of conventional–unconventional oil are revealed.

2 The Songliao Basin is a huge interior lake basin formed in the Early Cretaceous under the control of the subduction and retreat of the western Pacific plate and the massive horizontal displacement of the Tanlu Fault Zone in Northeast China.

3 During the deposition of the Qingshankou Formation, strong terrestrial hydrological cycle led to the lake level rise of the ancient Songliao Basin and the input of a large amount of nutrients, resulting in planktonic bacteria and algae flourish.

4 Intermittent seawater intrusion events promoted the formation of salinization stratification and anoxic environment in the lake, which were beneficial to the enrichment of organic matters.

5 Biomarkers analysis confirms that the biogenic organic matter of planktonic bacteria and algae modified by microorganisms plays an important role in the formation of high-quality source rocks with high oil generation capability.

6 There are four favorable conditions for the enrichment of light shale oil in the Qingshankou Formation of the Gulong Sag, Songliao Basin: the moderate organic matter abundance and high oil potential provide sufficient material basis for oil enrichment; high degree of thermal evolution makes shale oil have high GOR and good mobility; low hydrocarbon expulsion efficiency leads to a high content of retained hydrocarbons in the source rock; and the confinement effect of intra-layer cement in the high maturity stage induces the efficient accumulation of light shale oil.

7 The restoration of hydrocarbon accumulation process suggests that liquid hydrocarbons generated in the early (low–medium maturity) stage of the Qingshankou Formation source rocks accumulated in placanticline and slope after long-distance secondary migration, forming high-quality conventional and tight oil reservoirs.

8 Light oil generated in the late (medium–high maturity) stage accumulated in situ, forming about 15 billion tons of Gulong shale oil resources, which finally enabled the orderly distribution of conventional–unconventional oils that are contiguous horizontally and superposed vertically within the basin, showing a complete pattern of "whole petroleum system" with conventional oil, tight oil and shale oil in sequence.

**Fig. 2.** A figure of shale gas knowledge annotation process, showing the annotated entities.

## 2.3 Knowledge Fusion of Shale Gas Database

There are many data sources in the shale gas knowledge base, including experimental analysis data and reports, papers, databases, and shale gas production data and research reports, in which the naming rules, units and languages may be inconsistent. In order to reduce the ambiguity of the knowledge base, we refer to the industrial shale gas

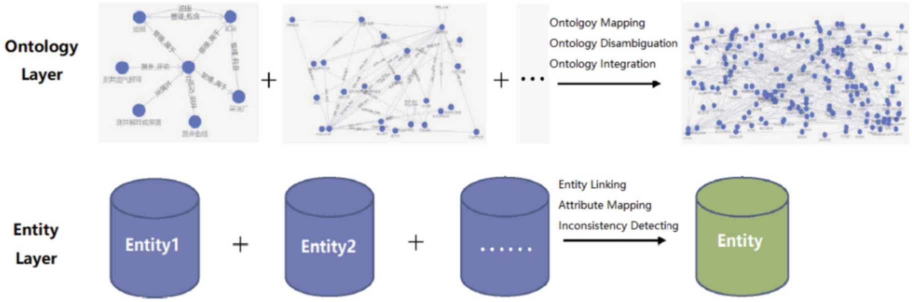
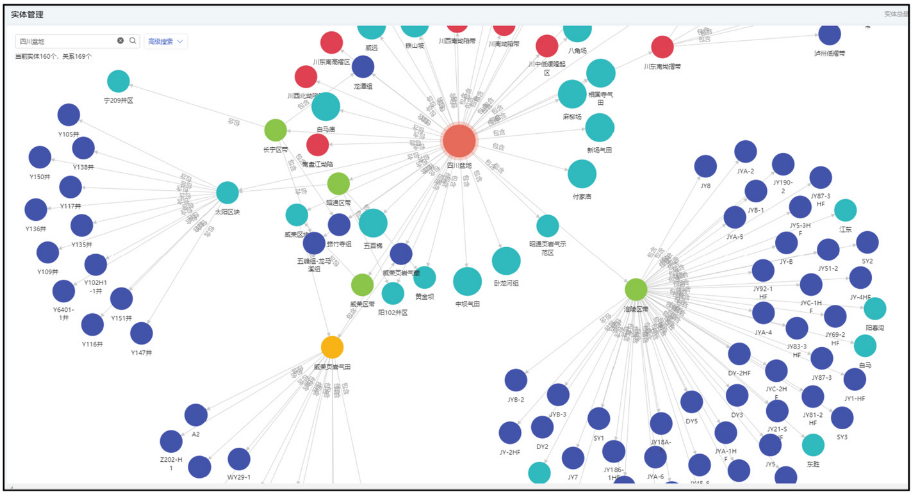


**Fig. 3.** General knowledge fusion process from both ontology and entity layers.

exploration and development data standards to realize the integration of knowledge extracted from various data sources. The process is called knowledge fusion. The fusion of shale gas knowledge includes the fusion of ontology layer and physical layer (see Fig. 3). The fusion of concept layer is mainly based on the knowledge expansion of shale gas recognition ontology, and the fusion of physical layer adopts entity linking technology. Firstly, based on the shale gas system knowledge formed by knowledge system classification, candidate entities are selected from various data sources through search engines. Subsequently, the supervised learning method is applied to train the candidate entity ranking model by manually labeling the training set, and the candidate entities are sorted. Finally, the fusion of the entity layer is completed through the entity similarity algorithm [6]. Knowledge fusion results need to be reviewed manually before entering the shale gas knowledge base.

After the steps of ontology construction, knowledge extraction and knowledge fusion, the shale gas knowledge base can be constructed. Compared with traditional knowledge bases, shale gas professionals can easily retrieve and view shale gas-related entities, attributes and relationships, and analyze them through visual and user-friendly interface (see Fig. 4).


实体管理

四川盆地

高级搜索

当前实体160个，关系169个

**Fig. 4.** A diagram of shale gas knowledge database with an interactive interface.

## 3 Application of the Shale Gas Knowledge Base

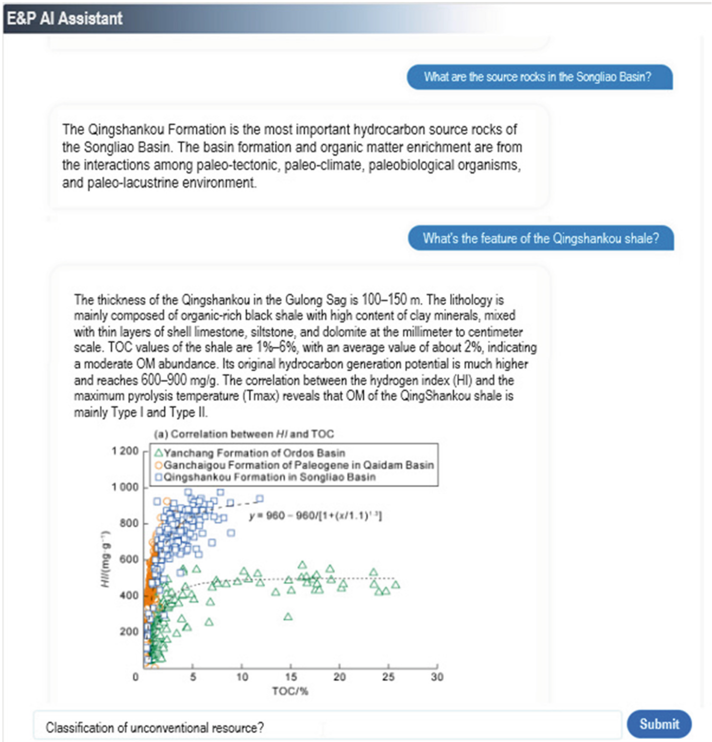
After the construction of shale gas knowledge base, various intelligent applications such as knowledge question-answering, search and recommendation systems can be implemented. Knowledge base question answering (KBQA) is one of the typical applications of knowledge base. It analyzes natural language problems, uses the knowledge base for querying and reasoning, and finally generates answers (see Fig. 5). Comparing with the current popular ChatGPT's generative question answering system, the answers from the KBQA system are more accurate since they come from the knowledge base. More importantly, there is no ChatGPT's incorrect or nonsensical answers to unknown questions. Reliability is usually the cornerstone of industrial question answering systems, so a knowledge base can verify and supplement the answers generated by ChatGPT.

Based on the knowledge base, graph machine learning can also be performed. Graph machine learning is similar to machine learning and can be used for tasks such as node classification, relationship prediction, clustering, and regression [8]. However, conventional machine learning and deep learning algorithms are not suitable for graph data, so graph representation learning can be used to convert graphs into feature vectors for traditional machine learning such as regression and classification. Common graph representation learning algorithms include node2vec and GCN. This research takes shale gas wells as the research object, and uses the feature engineering method to query the attributes and related nodes of shale gas wells, forming the data of the subsequent machine learning algorithm. Taking shale gas wells as the starting node, this paper sorts out 424 shale gas wells with 32 attributes available in the shale gas knowledge base. The Pearson correlation coefficient algorithm is used to generate the thermal diagram of shale gas well productivity factors (Fig. 6). From the figure, it can see that the EUR of shale gas wells is strongly correlated with the first-year production rate, the testing production, fracturing section length, lateral length, and fracturing fluid volume. The correlation coefficients on the five factors reach 0.91, 0.82, 0.54, 0.49 and 0.48, respectively.

### 3.1 EUR Predictions of Shale Gas Well

Based on the shale gas knowledge base, this study uses automated machine learning algorithms to predict the EUR of shale gas wells. The real EUR values were calculated by RTA (rate transient analysis), which is a modern tool to better understand the production and reserves of a reservoir or a well. Automated Machine Learning (Auto-ML) has become a trending topic in industry and academic artificial intelligence (AI) research in recent years. AutoML shows great promise in providing solutions for AI in regulated industries in providing explainable and reproducible results. AutoML allows for greater access to AI development for those without the theoretical background currently needed for role in data science. Every step in the current prototypical data science pipeline, such as data preprocessing, feature engineering, and hyperparameter optimization, can done automatically by AutoML.

In this study, we employ AutoML frameworks such as AutoGluon, TPOT, H2O, and Auto-Sklearn to conduct EUR prediction. Using AutoML for prediction tasks is a common approach to streamline and automate the machine learning pipeline. These frameworks provide automated solutions for various stages of the machine learning



What are the source rocks in the Songliao Basin?

The Qingshankou Formation is the most important hydrocarbon source rocks of the Songliao Basin. The basin formation and organic matter enrichment are from the interactions among paleo-tectonic, paleo-climate, paleobiological organisms, and paleo-lacustrine environment.

What's the feature of the Qingshankou shale?

The thickness of the Qingshankou in the Gulong Sag is 100–150 m. The lithology is mainly composed of organic-rich black shale with high content of clay minerals, mixed with thin layers of shell limestone, siltstone, and dolomite at the millimeter to centimeter scale. TOC values of the shale are 1%–6%, with an average value of about 2%, indicating a moderate OM abundance. Its original hydrocarbon generation potential is much higher and reaches 600–900 mg/g. The correlation between the hydrogen index (HI) and the maximum pyrolysis temperature (Tmax) reveals that OM of the QingShankou shale is mainly Type I and Type II.

(a) Correlation between *HI* and TOC
△Yanchang Formation of Ordos Basin
○Ganchaigou Formation of Paleogene in Qaidam Basin
□Qingshankou Formation in Songliao Basin
$y = 960 - 960/[1+(x/1.1)^{1.3}]$
*HI*/(mg·g⁻¹)
1 200
1 000
800
600
400
200
0
5
10
15
20
25
30
TOC/%

Classification of unconventional resource?

**Fig. 5.** Question-answering system based on shale gas knowledge base, showing the top 20 countries with potential for unconventional oil recoverable resources.

process, including feature engineering, model selection, hyperparameter optimization, and model evaluation.

The AutoGluon is an open-source AutoML framework from Amazon, which requires only a single line of Python to train highly accurate machine learning models on an unprocessed tabular dataset [10]. TPOT (Tree-based Pipeline Optimization Tool) is a popular open-source Auto-ML library in Python. It uses genetic programming to evolve an optimal pipeline of pre-processing techniques and machine learning models [11]. H2O is an open source, in-memory, and scalable machine learning and predictive analytics platform that allows users to build machine learning models on big data and provides easy productionalization of those models in an enterprise environment [7]. Auto-Sklearn is an open source AutoML toolkit, which includes latest research on automatically configuring the AutoML system itself and contains a multitude of improvements which speed up the fitting the AutoML system. It automatically sets the Model selection, decides whether it can use the efficient bandit strategy Successive Halving and uses meta-feature free Portfolios for efficient meta-learning.

The production data of 424 shale gas wells are split into training and testing datasets by a ratio of 0.85:0.15, which means there are 64 wells in the testing dataset. The training data of 360 wells are used to train the AutoML models. The performances of

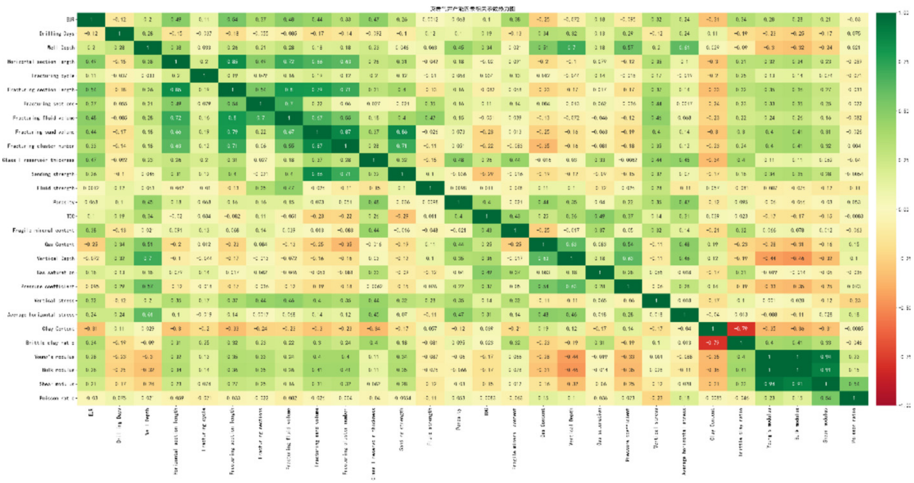


**Fig. 6.** Thermal diagram of correlation coefficient of shale gas well productivity factors

these AutoML frameworks for the testing dataset are listed in the table below, which shows the evaluation metrics of mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), Mean Absolute Percentage Error (MAPE), and the best model from each framework. The best model from TPOT, Auto-Sklearn and H2O is ExtraTreesRegressor. The study chooses the TPOT model as the best suitable algorithm for EUR prediction regression task after comprehensive comparison.

**Table 1.** The performance table of the four AutoML frameworks.

| AutoML Framework | MSE | RMSE | MAE | MAPE | Best Model |
|---|---|---|---|---|---|
| TPOT | 0.0604 | 0.2459 | 0.1846 | 22.7571 | ExtraTreesRegressor |
| Auto-Sklearn | 0.0685 | 0.2617 | 0.1811 | 20.4362 | ExtraTreesRegressor |
| AutoGluon | 0.1078 | 0.3283 | 0.2454 | 33.6250 | StackedEnsemble |
| H2O | 0.0670 | 0.2589 | 0.2012 | 33.7756 | ExtraTreesRegressor |

## 4 Results

The ExtraTreesRegressor model from TPOT is employed to conduct training and testing on the shale gas well data. The ratio of training and testing data is 0.85:0.15. TPOT tries to fit the training data using a few algorithm models like LightGBM, RandomForest, NeuralNetFast and so on. Meanwhile, it interpolates the missing values in the shale gas well data and optimizes kinds of parameters for different models. The EURs calculated by RTA and predicted from TPOT on both training and testing data are shown in Fig. 7. The MSE and MAE on the testing data are 0.0604 and 0.1846, while the MSE and MAE on the training data are 0.007 and 0.06, which implies somehow overfitting.

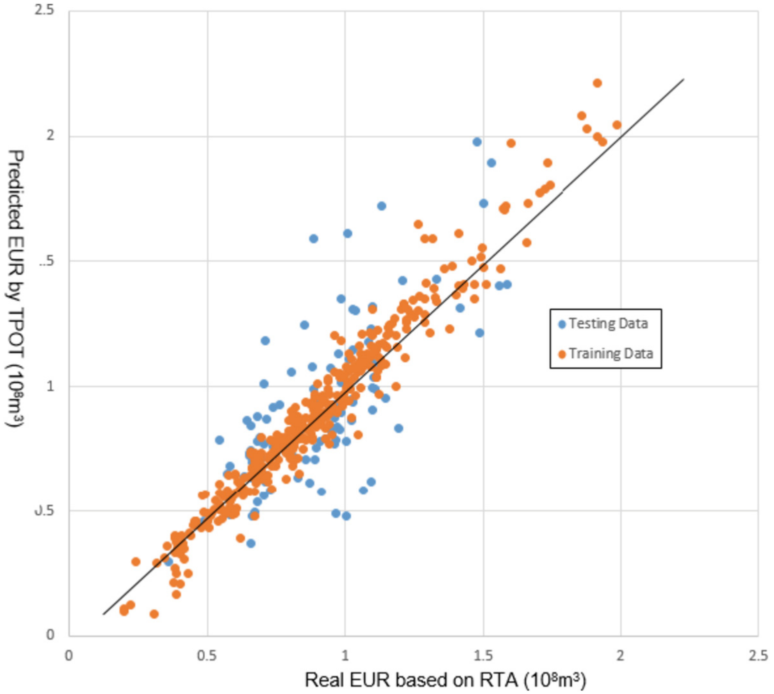


**Fig. 7.** EUR predictions of shale gas wells, showing the predictions on training and testing data

In addition, ExtraTreesRegressor can compute the importance of features to the EUR shale gas wells. A feature's importance score represents the performance drop that results when the model makes predictions on a perturbed copy of the data where this feature's values have been randomly shuffled across rows. A feature score of 0.01 would indicate that the predictive performance dropped by 0.01 when the feature was randomly shuffled. The higher the score a feature has, the more important it is to the model's performance. If a feature has a negative score, this means that the feature is likely harmful to the final model, and a model trained with the feature removed would be expected to achieve a better predictive performance. The top 5 most important factors to the EUR are high-quality bed thickness, fracturing section length, fracturing fluid volume, Fragile mineral content, Well Depth, respectively (see Fig. 8), which matches experts' experiences in shale gas well production very well.

In conclusion, the knowledge graph technology can be used to construct the database of shale gas exploration and production, which will be able to extract data and knowledge from various unstructured, semi-structured and structured data sources and then form a comprehensive shale gas knowledge base. Intelligent applications such as question-answering, search engine, EUR prediction can be subsequently built up based on the knowledge base. The procedure and experience on knowledge base construction and application by using cutting-edge technologies could be rolled out to other areas.

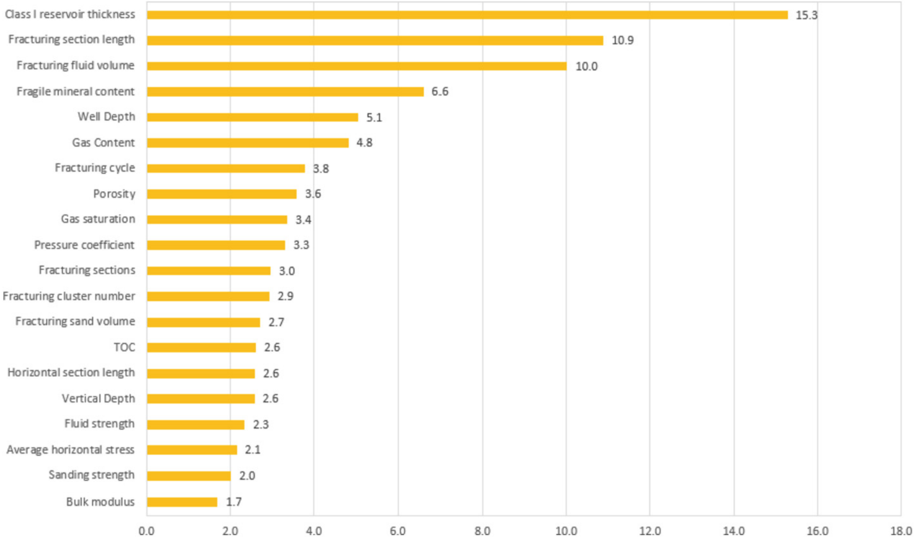


**Fig. 8.** Importance chart of attributes to the EUR of shale gas wells

## References

1. AutoGluon Homepage, https://auto.gluon.ai/stable/index.html. Accessed date: 2023/2/10
2. Auto-Sklearn Homepage, https://www.automl.org/automl/auto-sklearn/, last accessed 2023/5/13
3. Chaudhri, V. K., Chittar, N., Genesereth, M.: An Introduction to Knowledge Graphs. https://ai.stanford.edu/blog/introduction-to-knowledge-graphs/, last accessed 2023/3/12
4. Daixin, W., Peng, C., Wenwu, Z.: Structural Deep Network Embedding. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), pp. 1225–1234. Association for Computing Machinery, New York, NY, USA (2010)
5. Guangming, Z., Tianyuan, Y., Dongxi, T., Song, H.: Knowledge extraction of Chinese Ethnic Medicine based on BERT BiLSTM-CRF Model. Journal of Wuhan University (Science Edition) **67**(05), 393–402 (2021)
6. Guoqiang, L., et al.: Construction of Oil and Gas Reservoir Logging Knowledge Graph and Its Intelligent Identification Method. Pet. Explor. Dev. **49**(03), 502–512 (2022)
7. H2O Homepage, https://docs.h2o.ai/h2o/latest-stable/h2o-docs/welcome.html, last accessed 2023/5/10
8. Hamilton, W.: Graph Representation Learning, McGill University (2020)
9. Meiling, F., Lei, X., Ying, X.: An Entity Extraction Method for Electronic Target Atlas Based on BERT-BiLSTM-CRF Model. Journal of Aerospace Early Warning Research, 36 (03), 206–210+216 (2022)
10. Nick E., Jonas M., Alexander S., Hang Z., Pedro L., Mu L., Alexander S.: AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. arXiv preprint arXiv:2003.06505 (2020)

11. TPOT Homepage, https://github.com/EpistasisLab/tpot, last accessed 2023/5/15
12. Youwei, H., Zhiyue, H., Yong, T., Jiazheng, Q., Junjie, S., Yong, W.: Shale gas well production evaluation and prediction based on machine learning. Oil Drilling & Production Technology **43**(04), 518–524 (2021)

# Optimization of Data Insight Tool Based on Engineering Technology Data Governance Project in Ultra-deep Oil & Gas Fields

Qiang Zhang[1], Chun-lin Hu[2](✉), Rui Chen[3], Ke-cheng Jiang[2], Xin Li[2], Nan Xiao[3], Qing-gang Yang[2], and Bing-bing Zhou[2]

[1] PetroChina Tarim Oilfield Company, Xinjiang, China
[2] PetroChina Kunlun Digital Intelligence Technology Co., Ltd., Beijing, China
huchunlin39@cnpc.com.cn
[3] PetroChina Tarim Oilfield Company, R&D Center, Xinjiang, China

**Abstract.** Based on the engineering technology data in the ultra-deep oil & gas fields, this paper utilizes data insight tool to identify and extract information from various types of data stored in documents with text or tables, which meets the needs of data governance project. If the document information is about text content, the natural language processing (NLP) method is directly selected for recognition; If the document information is a table, it is necessary to convert the table into a heterogeneous data table with Date-Frame format first by Python language, and then recognize and extract it. These two processing methods can successfully convert unstructured data to structured data, solving the problem of low accuracy and low timeliness of extracting information from different documents. The NumPy & Pandas learning with Python language and other algorithms/functions play an important role in building metadata models, labeling fields, and training backend algorithms of data insight tool structure. The target trained extraction model is very crucial to the identification and extraction of various information. Relying on this and later, the qualified data generated after steps of extraction of target documents, selection of matching data for review and multi-level audit evaluation will be marked with "EDG", which is the main data source of various professional databases of Tari Oilfield and the guarantee of the capacity and

This paper was prepared for presentation at the 2022 International Field Exploration and Development Conference in Urumqi, China, 16–18 August, 2022.

This paper was selected for presentation by the IFEDC Committee following review of information contained in an abstract submitted by the author(s). Contents of the paper, as presented, have not been reviewed by the IFEDC Technical Team and are subject to correction by the author(s). The material does not necessarily reflect any position of the IFEDC Technical Committee its members. Papers presented at the Conference are subject to publication review by Professional Team of IFEDC Technical Committee. Electronic reproduction, distribution, or storage of any part of this paper for commercial purposes without the written consent of IFEDC Organizing Committee is prohibited. Permission to reproduce in print is restricted to an abstract of not more than 300 words; illustrations may not be copied. The abstract must contain conspicuous acknowledgment of IFEDC. Contact email: paper@ifedc.org.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 55–64, 2024.
https://doi.org/10.1007/978-981-97-0272-5_5

quality of the data lake. Examples show that the data insight tool has strong adaptability, obvious optimization effects, and superior performance compared to other extraction tools. The development and application of data insight tool have significantly improved the identification and extraction ability for engineering data of ultra-deep oil & gas fields, improved the identification accuracy and extraction speed, and met the needs of data governance.

**Keywords:** Data insight tool · Information extraction and metadata

## 1 Introduction

With the development of computer technologies such as cloud computing and big data, people from all works of life are gradually realizing the value of data. In the process of ultra-deep oil & gas exploration and production, there is also a large amount of data, which is generated in different ways at different times and stored in the form of documents. There are not only geophysical, drilling and logging data, but also gas testing, geological structure, and downhole operation data, as well as scientific research report data. For documents from the same data source, there are also storage differences such as text, horizontal tables, vertical tables, and two-dimensional tables. Overall, it appears as a long collection and survival time, discontinuous, and mixed with other data, diverse storage types, complex document structure, and large and disorderly data volume.

In order to make full use of these data, it is necessary to treat them as a whole and convert all unstructured data into structured data, so as better serve the engineering technology data governance project of ultra-deep oil & gas fields, which has practical significance [1, 2].

Unstructured data refers to data with irregular or incomplete structure, no predefined data model, and inconvenient to be represented by the two-dimensional logical tables of the database, including office documents, text, images, HTML, reports, images, and audio/video information in various formats, while structured data, also known as row data, refers to data logically expressed and realized through the two-dimensional table structure. It strictly follows data format and length specifications, mainly utilizing relational databases for storage and management.

The conventional data structuring is achieved by constructing an information extraction model. The specific method is to manually mark the information to be extracted in the sample to obtain the training sample set, then select the appropriate supervised learning algorithm, train the model with the training sample set, and finally obtain the model for information extraction.

This method may not always be effective. For example, the sample and text to be extracted are in a fixed format, such as ID cards and invoices. When the format of the text to be extracted is very consistent with the sample, this method can obtain a model with high extraction accuracy. When there are differences in format between, the accuracy of model extraction decreases, and the larger the difference, the lower the accuracy. Although the accuracy of the model can be improved by increasing the number of samples, but an increase in the number of samples often means an increase in training difficulty. In the oil and gas industry, the processes of oil and gas exploration and production are complex, and the documents generated by different periods/methods may have significant differences. Even increasing the number of samples cannot achieve the accuracy value required for practical applications [3].

Data insight is a method that can efficiently extract information, the tool equipped with corresponding devices and storage media, especially suitable for environments where there are differences in the format of samples and documents to be extracted, significantly improving the accuracy of information extraction models.

## 2 Features of Data Document in Ultra-deep Oil & Gas Fields

At the initial stage of data insight tool development, the unstructured data of engineering technology in ultra-deep oil & gas fields was briefly classified according to the type of data documents, and the following rules were summarized:

(1) The same type of data documents of different wells in different oilfields may have large differences, and similarly, documents of different data types also have different structures; (2) Within the same document, there may be both text and table content, and table documents of the same type may also contain horizontal, vertical, or two-dimensional tables; (3) The probability that data simultaneously exists in three types of text, table, and image in a data document is high. Among the total data of all types of documents, table data accounts for the largest proportion, reaching over 90%. The image data is about 8%, and text data less than 2%. (4) Table documents mainly include basic table, cross page table, two-dimensional table, and transposition table. Among them, cross page table accounts for the largest proportion, reaching over 60% of the total table documents. Transposition table is about 15%, while basic tables and two-dimensional tables are about 20% and 5%, respectively.

In response to the features of data document in the ultra-deep oil & gas fields, the key task of the data insight tool primarily solves the problem of extracting information from table data documents, followed by image and text information extraction.

## 3 Integral Development of the Data Insight Tool

The data insight tool typically consists of two parts: hardware and software. The hardware includes mechanical and electronic apparatus /devices or storage media used to support data conversion, storage, transmission, and communication responses such as computers, circuit boards, memories, and transceivers; The software includes corresponding networks, algorithms, sample sets, training models, databases, and data communication and control.

### 3.1 Hardware Composition and Network Architecture of the Data Insight Tool

Specifically, the data insight tool includes three functional blocks: data processing center, database, and user terminal. Each function block contains hardware and software to complete all functions one by one, that is, trigger user information extraction requests, generate metadata and target documents, collect and call training sample sets, generate

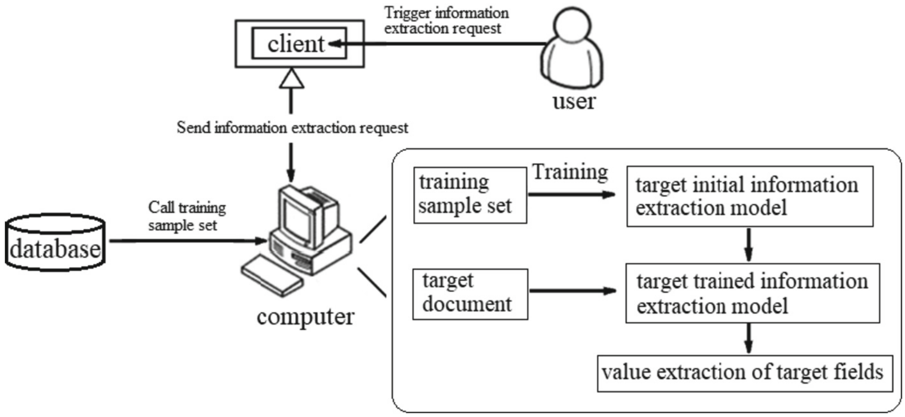

client
Trigger information extraction request
user
Send information extraction request
Call training sample set
database
computer
training sample set
Training
target initial information extraction model
target document
target trained information extraction model
value extraction of target fields

**Fig. 1.** Hardware composition and network architecture diagram of the data insight tool

information extraction models for target training, and extract the values of target fields. The hardware composition and network architecture diagram shown in Fig. 1.

The user terminal (shown as client in Fig. 1) is the port where users make requests and receive responses, and is the response center for data processing center exchange with the outside world. It is equipped with information extraction application software, which is used to receive information extraction requests triggered by users. Users can access the operation interface of the client or enter the website to extract web pages, sending information extraction requests to the data processing center.

The database (shown as database in Fig. 1) is the storage center of structured data, target documents and training samples/training sample sets. The data processing center calls the training samples from the database to form a training sample set according to the user's information extraction request.

The data processing center (shown as computer in Fig. 1) is the center for data reception and transmission, algorithm collection, training models, and communication control. It mainly includes various algorithms for constructing metadata models, training sample sets for completing data recognition and extraction, and executing instructions. It is the most core part of the data insight architecture.

The data processing center adopts a modular design concept and is divided into modules such as reception, training, determination, conversion, and acquisition. Briefly, the receiving module receives extraction requests to form initial metadata and target documents; The determination module obtains the training sample set corresponding to the initial metadata; The training module generates a target initial information extraction model and trains it to generate a target trained extraction model corresponding to the initial metadata model; The conversion module is used to determine and convert the file format of training samples; The acquisition module extracts target fields from the target document using an information extraction model based on target training.

### 3.2 Procedure of Information Extraction

The procedure of information extraction for data insight sequentially involves the following steps:

(1) Receive user triggered information extraction requests, generate target fields to be extracted based on their needs and selections, record the target documents to which these target fields belong, their document types, and document locations, and construct an initial metadata model.
(2) Select training sample documents that meet the conditions and convert their format into a unified and recognizable format. Collect them together to form a training sample set, and label the target fields and field values of all documents in the sample set.
(3) Referring the metadata model and initial information of the target to build a target initial extraction model, and configure the parameters in the model.
(4) Extract the labeled target fields and field values from each training sample in the training sample set, and use them to train the target initial extraction model.
(5) Set the threshold for extraction recall and extraction accuracy. If the target initial extraction model does not reach the threshold for extraction recall or extraction accuracy, adjust the model parameters and keep training with the sample set until these two thresholds are exceeded.
(6) Now the target initial extraction model can be considered as a target trained information extraction model and shines upon the corresponding metadata model.
(7) Using the target trained information extraction model to extract the target fields in the target document, the values of the target fields can be obtained.

## 4 Data Insight Analysis and Strategies

In order to accurately and quickly complete information extraction, specific analysis and testing were conducted on each process and step, and corresponding strategies were summarized.

**Metadata and Training Sample Set Selection Strategy**
A few of metadata and corresponding training sample sets are preconfigured in the database, so that documents of different structural types correspond to different metadata. When a user requests information extraction, appropriate metadata and training sample sets are filtered out, and the types of samples are then distinguished. During the training process of the target trained information extraction model, there is no need to filter the training samples anymore, just learn all the samples in the selected training sample set. Therefore, the target trained extraction model can have a high accuracy, and requires a small number of samples, resulting in fast extraction speed.

**Format Transformation Strategy for Training Sample Documents**
Check the file header flag to clarify the format of the training sample. If the sample is in PDF format, use common format conversion tools to directly convert the sample format to Excel format; If it is a Word document, first convert the format of the Word document to PDF format, and then convert it to Excel format. This strategy can unify

the document format of training samples and ensure that the content in the document will not be disordered after format conversion. Meanwhile, compared to other methods such as converting each page of the sample document into images and then performing image recognition, format conversion can improve the speed of model training and usage, especially when the number of pages in the document is large, the speed improvement is more significant.

**Extraction Strategies for Text and Table Documents**

Using NumPy and Pandas in Python language, convert all data in the document into Date-Frame format data, and separate the text content and table content in each training sample by presetting the data length of heterogeneous data tables.

Various algorithms are set in the target initial information extraction model, such as N-gram language model for table name query, cosine similarity algorithm for obtaining similar table names, OCR for obtaining information in pictures, named entity recognition, short-term memory LSTM, conditional random field CRF and Glove algorithm, etc. for DataGrid Cell segmentation.

If it is text content, use natural language processing NLP to extract text content. Sequentially perform extraction methods and strategies such as word segmentation, word embedding, and named entity recognition, mark the extraction relationship between the target field and its field values, and extract the feature values of the target field. If it is a table content, use the concatenation function to concatenate the row data in the worksheet into a string, use the judgment function to determine whether the tables in each worksheet are continuation tables, and use the preset deletion function to delete unmarked irrelevant items and items without data in the table.

These algorithms and functions have improved the efficiency of learning and training, reducing the probability of errors.

**Difference Sample Determination and Strategy**

When there are difference samples in a training sample set, specific analysis pointing at the target fields of the difference samples is necessary.

If the difference is only about in the text feature values of the difference target field, it indicates that an error may have occurred during the text label process, and the difference target field in the difference sample needs to be discarded. If the differences happened both in position and contextual feature values, it indicates that the position of the target field in the difference sample may be different from that in the other samples. The feature values of the difference target field can be stored as the second feature values of the target field in the target initial information extraction model, and keep going on extracting target field and its field values from the training sample. This strategy can store multiple feature values for the same target field, and then calculate the feature value by the way of weight average or linear regression, which has a significant impact on sample reinforcement. If the differences simultaneously appear in position, text, and contextual feature values, it indicates that there may be errors in the selection of training samples and the difference sample need to be discarded.

These strategies above could ensure that during the training process, the initial target information extraction model has fewer training samples and the most types, making it more accurate and faster.

## 5 Application Examples

In order to verify the advantages of data insight tool in governance time, the data insight tool and other popular extraction tools were selected to synchronously extract logging data of ultra deep oil fields with “inclination” as the target field. The comparison of extraction effects is shown in Fig. 2 and Fig. 3. 50 Wells are selected in Fig. 2, while 500 Wells selected in Fig. 3 to demonstrate the extraction capability in more details. Here are three extraction tools, namely: Data Insight Extraction in the first row, Baidu Table Analysis Extraction in the second row, and Manual Extraction in the third row. In Fig. 2 and Fig. 3, blue represents the time for labeling samples, orange represents the training time, and gray represents the extraction (governance) time. The horizontally distributed numbers “0”, “200”, “400”, “600”, “800”, as well as “0”, “2000”, “4000”, and “6000” in the figures represent the time expenditure for governance, in minutes. Their specific values are related to the size of the “Inclination” data.

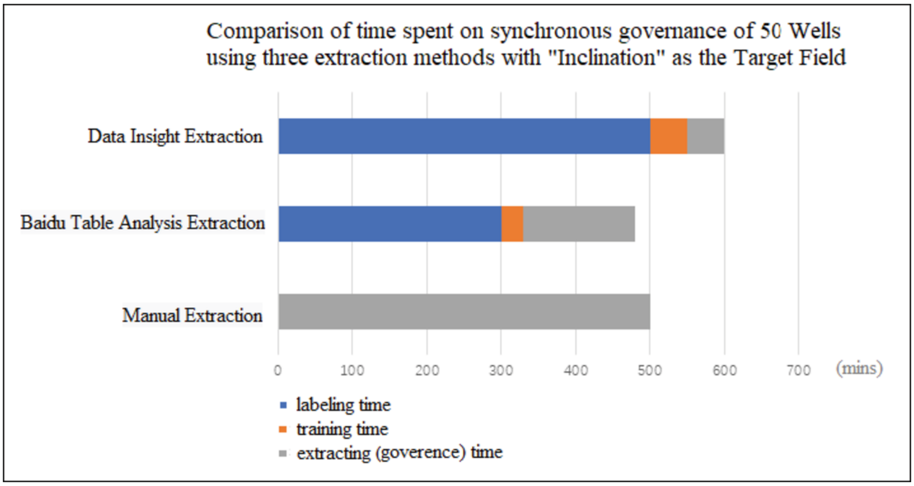


**Fig. 2.** The data insight tool and other popular extraction tools were selected to synchronously governance the logging data of 50 Wells in ultra deep oil fields, with “Inclination” as a target field

From Fig. 2 we can see that when the number of Wells is small, such as less than 50 wells, the Manual Extraction is the fastest method due to the time expenditure for data insights and other extraction tools to label samples. As the number of Wells increases, such as 500 wells shown in Fig. 3, the advantages of artificial intelligence are reflected. In Fig. 3, the governance time using Data Insight Extraction is the shortest, only half of the time used by Baidu Table Analysis Extraction, and 1/5–1/6 of the time used by Manual Extraction.

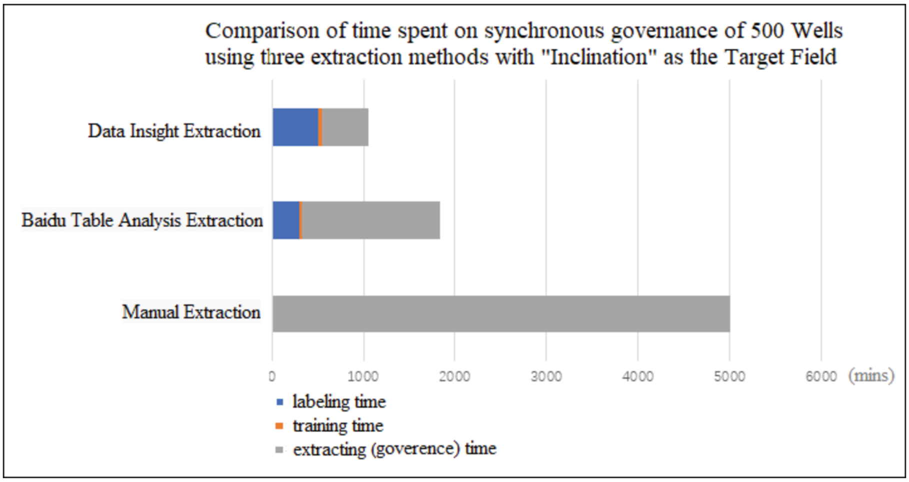


**Fig. 3.** The data insight tool and other popular extraction tools were selected to synchronously governance the logging data of 500 Wells in ultra deep oil fields, with "Inclination" as a target field.

It can be foreseen that for the huge amount of data in the ultra-deep oil fields, data insight tool will have significant advantages in governance time.

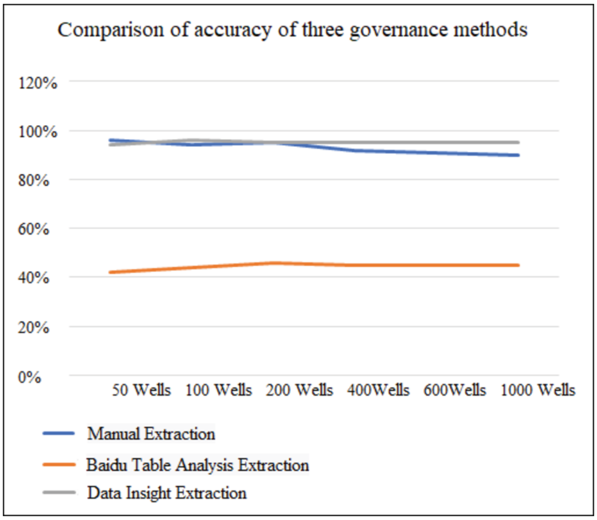


**Fig. 4.** The data insight tool and other popular extraction tools were selected to synchronously governance the same number of Wells of the logging data in ultra deep oil fields, with "Inclination" as a target field.

For the sake of verifying the advantages of data insight tool in governance accuracy, data insight tool and other popular extraction tools were also selected to synchronously extract data in the same number of Wells in the logging data of ultra deep oil fields with "Inclination" as a target field. The comparison of extraction effects is shown in Fig. 4. Here the blue line represents extraction accuracy of Manual Extraction, orange line represents extraction accuracy of Baidu Table Analysis Extraction, and gray line represents the accuracy of Data Insight Extraction. The horizontal characters "50 wells", "100 wells", "200 wells", "400 wells", "600 wells", and "1000 wells" in the figure represent the number of governance Wells, while the vertical characters represent a sketch map of accuracy percentage. Likewise, their specific values are related to the size of the "Inclination" data.

From Fig. 4 we can see that both Data Insight Extraction and Manual Extraction have an ideal information extraction accuracy (as shown by the gray and blue lines) of over 90%, regardless of whether the number of Wells is small (such as 50 Wells) or large (such as 1000 Wells). However, as the number of Wells, meaning the amount of governance data increases, the extraction accuracy of Manual Extraction (as shown by the blue line) slightly decreases. While using Baidu Table Analysis Extraction, regardless the number of governance Wells large or small, the extraction accuracy (orange line) is below 50%.

The above examples indicate that the data insight tool has high extraction accuracy and fast extraction speed, and their information extraction ability far exceeds that of other common products on the market, meeting the audit requirements of business departments.

## 6 Conclusion

(1) The metadata of data insight tools covers a wide range, including different target fields that will be extracted, different documents storing these target fields, and their respective types and locations. These documents are all technical data generated during the ultradeep oil & gas exploration and production. When users trigger information extraction requests, multiple metadata can be generated simultaneously for backup. The target trained model is trained using a training sample set, and the type of samples in the training sample set is the same as that of the target document. All training samples are labeled with the target fields and their field values. Therefore, there is not much difference between the format of the target document and the format of the samples, and a high accuracy information extraction model can be quickly obtained.

(2) The data insight tool converts unstructured data into structured data by means of establishing metadata models, marking fields, background algorithm training etc., which plays a key role in the construction of various professional databases in ultra deep oil & gas fields. Qualified data generated after the steps such as extracting target documents, selecting matching data for review and multi-level audit and evaluation will be marked with "EDG", which is the main source of the database of ultra deep oil & gas fields and the guarantee of the capacity and quality of the data lake [4].

(3) The example verification shows that both the accuracy and speed of data insight tool used for information extraction remain at a high level, which can meet the audit requirements of business departments. It has strong adaptability, obvious optimization effects, and superior performance, far surpassing other common extraction tools in the market.

(4) From design to development, the data insight tool absolutely conforms to the current situation of technical data of ultra deep oil fields, and its accuracy and governance speed meet the requirements of practical application. It is of great significance to promote the implementation of engineering technology data governance project of Ultra Deep Oil & Gas Fields.

## References

1. Gao, H., Li, J., Cao, Y., et al.: Construction of expeditor seismic data management system in Tarim oilfield. Petrol. Geophys. Explor. **43**(supplement1), 182–196 (2008)
2. Lu, Z., Chen, R., et al.: Discussion on real time data quality management of drilling engineering in Tarim Oil Field **31**(3), 118–121 (2020)
3. Li, X.: Tarim oilfield surface engineering database system. Oil Gas Field Surface Eng. **21**(2), 118 (2002)
4. Luo, C., Tang, Y., et al.: Application and prospect of collaborative research on E&P dream cloud in Tarim oilfield. China Petrol. Explor. **25**(5), 50–55 (2020)

# Intelligent Evaluation Method of Cement Bond Quality Based on Convolutional Neural Network

Xiang Wang[1](✉), Hui Ding[2], Gang Yu[2], Rui Liu[2], and Zheng-chao Zhao[2]

[1] School of Petroleum and Natural Gas Engineering, Changzhou University, Changzhou, China
xiangwang@cczu.edu.cn

[2] Tarim Oilfield Company, PetroChina, Korla, China

**Abstract.** The quality of cement bond is related to the safety of oil and gas well production and the service life of casing. At present, acoustic variable density logging (VDL) is the most widely used method for evaluating cementing quality in oil fields. The data interpretation of VDL still needs to rely on manpower, and the accuracy of interpretation results is restricted by human factors, and the workload is heavy. Oilfields have accumulated a large number of practically verified VDL interpretation results. It is of great research value and application potential to sort out these historical data and mine them with the help of deep learning technology, and establish an intelligent analysis method instead of humans to explain the cementing quality. In this study, the VDL cementing quality evaluation reports of several oil wells were collected. Through data preprocessing, the acoustic variable density images were standardized and segmented along the borehole direction. The cementation conditions of the first interface and the second interface corresponding to each segment of the acoustic variable density image were marked, and a sample set for cement bond quality evaluation was established. The cementing quality evaluation problem is transformed into an image classification problem, and the convolutional neural network method is introduced. On the basis of LeNet5, AlexNet and other classic image recognition architectures, considering the characteristics of acoustic variable density images, a personalized convolutional neural network (CBQNet) for cementing quality evaluation is designed, including 28 layers and more than 32 million learnable parameters.

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20-22 September 2023.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 65–77, 2024.
https://doi.org/10.1007/978-981-97-0272-5_6

Using historical cementing quality evaluation samples to train and analyze the performance of convolutional neural network, the results show that: CBQNet has a training accuracy rate of 95.9% and a verification accuracy rate of 95.4% in the first interface cementing quality evaluation. In the cementing quality evaluation of the second interface, the training accuracy rate reached 90.8%, and the verification accuracy rate reached 88.1%. It shows that the convolutional neural network realizes efficient and accurate interpretation of cementing quality by mining and learning the interpretation results of historical VDL data, and provides a new method for cementing quality evaluation.

**Keywords:** Cement Bond Quality Evaluation · Convolutional Neural Network · VDL Logging · Pattern Recognition

## Nomenclature

| | |
|---|---|
| $a_k$ | Output of the *k*th sample; |
| $C$ | Cross-entropy loss function; |
| $i, j$ | Number of neurons; |
| $j$ | Total number of neurons of layer; |
| $k$ | Number of sample; |
| $n$ | Total number of samples; |
| $relu_i(\boldsymbol{x})$ | ReLU function; |
| $softmax_i(\boldsymbol{x})$ | Softmax function; |
| $\boldsymbol{x}$ | Vector of parameters for each neuron in a neural network layer; |
| $x_i, x_j$ | Parameter for the *i*th and *j*th neurons; |
| $y$ | Label value; |
| $y_k$ | Label value of the *k*th sample |

## 1 Introduction

Cementing is a key technology in the process of oil and gas field development, and the quality of cement bond has a direct impact on the life and productivity of the well. During the cementing process, it is difficult to ensure good cementing quality in the entire well section due to the properties of the medium in the well, the cementing operation environment and various factors during the construction process [1–3]. Unqualified cementing quality may lead to reduced well life and oil layer pollution. How to evaluate the cementing quality reasonably, locate unqualified well sections in time, and give reasonable remedial measures has become an important task of cementing.

Cementing quality evaluation is mainly based on the analysis of the cementation of the two interfaces. Interface I is the cemented interface between the casing and the cement sheath, and interface II is the cemented interface between the cement sheath and the formation. No matter whether the cementation quality of the interface I or the interface II does not meet the standard, it is easy to cause downhole oil-water channeling, and even destroy the regional geostress balance, resulting in casing damage [4–6].

In the 1970s, acoustic variable density logging (VDL) technology was developed for cement bond evaluation. The VDL logging tool adopts the single-send and double-receive mode. The sound wave is emitted from the transmitter, and the sound wave passes through various interfaces in the well and is finally received by the receiver. There are two source distances for receivers. The 3ft source distance receiver is used to measure the casing wave sound amplitude, which is used for the evaluation of the interface I. The 5ft source distance receiver is used to measure the full wave of the sound wave, and the acoustic variable density image is obtained after processing, which can reflect the cement bonding of the interface I and interface II. VDL logging technology has become more and more mature and has become the most widely used cementing quality evaluation technology. However, at present, the data interpretation of VDL still needs to rely on manpower, and the accuracy of interpretation results is restricted by human factors, and the workload is heavy [7–9].

Big data and deep learning technology are causing a new round of technological revolution. Breakthroughs have been made in many fields such as image recognition, voice processing, and unmanned driving [10]. Petroleum companies are also actively introducing artificial intelligence technology to promote intelligent transformation and upgrading [11, 12]. At present, the oil field has accumulated a large number of practically verified VDL interpretation results. It is of great research to sort out these historical data and mine them with the help of deep learning technology, so that it can replace humans in cementing quality interpretation. This has great research value and application potential.

In this study, we propose to apply convolutional neural network to the problem of cementing quality evaluation. Firstly, a sample set of cementing quality evaluation will be prepared based on the historical VDL data and the corresponding cementing quality interpretation results. After that, a cement bond quality evaluation model will be established based on convolutional neural network. The sample set will be mined and learned, and the performance of the model will be analyzed.

The paper is structured as follows: Sect. 2 provides an description of the preparation process for the cementing quality evaluation sample set. Section 3 discusses the design concept and outcomes of the convolutional neural network architecture for cementing quality evaluation. Section 4 presents the training process and performance analysis results of the neural network for cementing quality evaluation. Finally, Sect. 5 concludes the paper.

## 2 Preparation of Cement Bond Quality Evaluation Sample Set

### 2.1 VDL Logging Interpretation Image

VDL is a commonly used cementing quality detection method in the field. The principle is to reflect the bonding quality between cement and casing, and between casing and formation by using the large difference in acoustic impedance between cement and mud (or water) on the attenuation of sound waves propagating along the axial direction of the casing [2].

The VDL tool adopts the single-send and double-receive mode. The sound wave is emitted from the transmitter, and the sound wave passes through various interfaces in the well and is finally received by the receiver. There are two source distances for receivers. The 3ft source-distance receiver measures the sound amplitude of the casing wave, which is used for the evaluation of the first interface of cementing. The receiver with a source distance of 5ft is used to measure the full wave train of the sound wave, and then the components and amplitude of the first arrival wave are extracted through data processing, and the sound wave variable density map is obtained, which can reflect the cement cementation of the interface I and interface II. There are black and white strips on the acoustic variable density image, and the intensity of the signal is represented by the color of the strips. In the acoustic variable density image, combined with geological information and cementing slurry information, the cementing quality analysis can be carried out according to the clarity of the full wave train strips [13].

Figure 1 illustrates a typical image of cementing quality interpretation results. The figure presents six types of logging information, namely natural gamma ray logging (GR), caliper logging (CAL), acoustic amplitude logging (AC), acoustic time difference logging (CBL), magnetic positioning logging (CCL), and acoustic variable density logging (VDL). Additionally, the image displays the cementing quality analysis results of two interfaces, namely interface I and interface II, on the left side. The different tiles in the image correspond to different cementing qualities, including five distinct interpretation results, namely good cementation, moderate cementation, poor cementation, mixed mud zone, and mud zone.

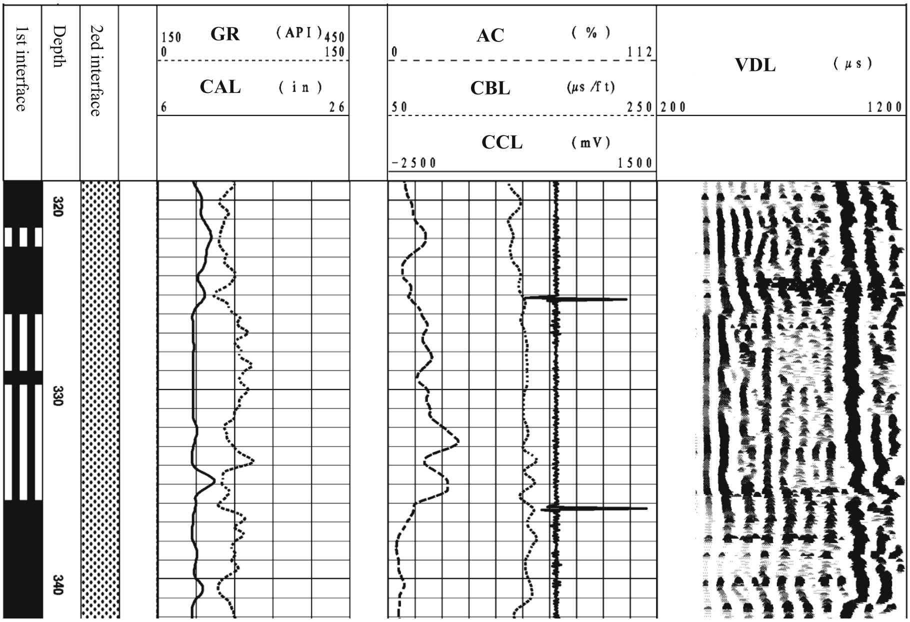
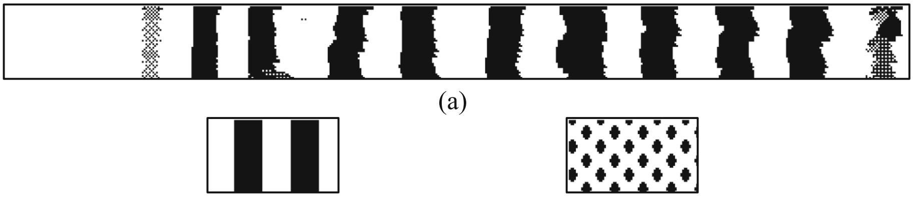


**Fig. 1.** A typical image of cementing quality interpretation results.

## 2.2 Sample Set Preparation

Cementing quality interpretation result images from oil fields were collected. The entire interpretation result image was segmented along the borehole direction with a width of 1m, results many independent images for each meter, each with a size of 1886 × 41 pixels.

Each meter of the interpretation result image was further cropped to intercept the VDL image part, the interpretation result part of the interface I, and the interpretation result part of the interface II. Specifically, the VDL image part was cropped to a size of 511 × 41 pixels, while the interpretation result parts of the first and second interface sections were cropped to sizes of 73 × 41 pixels each. An example of resulting images is presented in Fig. 2.



(a)

(b)

(c)

**Fig. 2.** An example of resulting images cropped from an interpretation result image. (a) VDL image part. (b) interface I interpretation result part. (c) interface II interpretation result part.

The VDL image in Fig. 2(a) is a black and white image that can be transformed into a matrix of size 511 × 41, where each element in the matrix takes a value of either 0 or 1. In this matrix, 0 represents a white pixel and 1 represents a black pixel. Similarly, the images in Fig. 2(b) and Fig. 2(c) are black and white and represent the interpretation results of the first and second interfaces. During the preparation of the sample set, the interpretation results of the wellbore interfaces were transformed into vectors using the one-hot encoding method, as shown in Table 1.

**Table 1.** Interpretation results and corresponding one-hot code.

| Image of interpretation results | Description of interpretation results | One-hot code |
|---|---|---|
| ■ | good cementation | [1 0 0 0 0] |
| ▥ | moderate cementation | [0 1 0 0 0] |
| ⁘ | poor cementation | [0 0 1 0 0] |
| ▦ | mixed mud zone | [0 0 0 1 0] |
| ▤ | mud zone | [0 0 0 0 1] |

The cement bond quality evaluation sample set was obtained by processing the interpreted image for each meter of each well. A total of 3351 samples were prepared in this study. Each sample contains an input image, and two labels representing the cementing quality of the first interface and the second interface, respectively.

## 3 Architecture Design of Convolutional Neural Network for Cementing Quality Evaluation

Given that the input for evaluating cement bond quality is the VDL image, a convolutional neural network (CNN) with robust image feature learning and classification abilities was chosen. CNN is the leading algorithm in computer vision research, especially in image recognition, and has demonstrated a range of successful applications [14]. As a deep learning algorithm, CNN is inspired by the visual cortex structure in animals that adaptively extracts spatial hierarchical information from images through layers of various visual neurons. CNN typically includes three kinds of layers, namely, convolutional layers, pooling layers, and fully connected layers. The convolutional and pooling layers are utilized for image feature extraction, where the former leverages different convolutional kernels to scan the feature maps for extracting features from diverse perspectives, and the latter reduces the dimensionality of the features. The fully connected layer maps the extracted features to the final output.

For different problems, the number and logical relationship of convolutional layer, pooling layer and fully connected layer are different, that is, the design of convolutional neural network architecture is different. Due to the inexplicability of neural network algorithms, the current neural network architecture design still lacks general standards and specifications, and relies more on experience and trial and error. According to the characteristics of the cementing quality evaluation problem, combined with the classic network architectures such as LeNet-5, AlexNet, VGGNet, GoogleNet and ResNet in the image recognition field [15], the architecture of the convolutional neural network for cementing quality evaluation is designed and named as CBQNet. Its architecture parameters are shown in Table 2.

The designed convolutional neural network, CBQNet, for evaluating the quality of cementing contains a total of 28 layers, including 6 convolutional layers that all utilize 3 × 3 small convolutional kernels and 3 pooling layers that all use 2 × 2 maximum pooling method. With the exception of the Softmax activation function used before the classification output, all intermediate layers use the ReLU activation function. The formulas of Softmax and ReLU activation functions are:

$$softmax_i(x) = \frac{\mathrm{e}^{x_i}}{\sum\limits_{j=1}^{J} \mathrm{e}^{x_j}} \ (i = 1, 2, 3, \cdots, J) \tag{1}$$

$$relu_i(x) = \max(0, \mathrm{e}^{x_i}) \ (i = 1, 2, 3, \cdots, J\,) \tag{2}$$

The CBQNet has a total of over 32 million learnable parameters. To avoid issues related to overfitting and lengthy training times, five dropout layers were added. During training, the dropout layers randomly select a certain proportion of neurons to stop

**Table 2.** Architecture parameters of CBQNet.

| Layer No. | Layer Type | Settings | Dimensions | Learnable Parameters |
|---|---|---|---|---|
| 1 | Image Input | | 511 × 41 × 1 | |
| 2 | Conv2D | Size: 3 × 3<br>No.: 32<br>Stride: 1 × 1 | 509 × 39 × 32 | Weights: 3 × 3 × 1 × 32<br>Bias: 1 × 1 × 32 |
| 3 | Activation | Function: ReLU | 509 × 39 × 32 | |
| 4 | Conv2D | Size: 3 × 3<br>No.: 32<br>Stride: 1 × 1 | 507 × 37 × 32 | Weights: 3 × 3 × 1 × 32<br>Bias: 1 × 1 × 32 |
| 5 | Activation | Function: ReLU | 507 × 37 × 32 | |
| 6 | Pooling | Type: Max Pooling<br>Size: 2 × 2<br>Stride: 2 × 2 | 254 × 19 × 32 | |
| 7 | Dropout | Ratio: 5% | 254 × 19 × 32 | |
| 8 | Conv2D | Size: 3 × 3<br>No.: 64<br>Stride: 1 × 1 | 252 × 17 × 64 | Weights: 3 × 3 × 32 × 64<br>Bias: 1 × 1 × 64 |
| 9 | Activation | Function: ReLU | 252 × 17 × 64 | |
| 10 | Conv2D | Size: 3 × 3<br>No.: 64<br>Stride: 1 × 1 | 250 × 15 × 64 | Weights: 3 × 3 × 64 × 64<br>Bias: 1 × 1 × 64 |
| 11 | Activation | Function: ReLU | 250 × 15 × 64 | |
| 12 | Pooling | Type: Max Pooling<br>Size: 2 × 2<br>Stride: 2 × 2 | 125 × 8 × 64 | |
| 13 | Dropout | Ratio: 5% | 125 × 8 × 64 | |
| 14 | Conv2D | Size: 3 × 3<br>No.: 128<br>Stride: 1 × 1 | 123 × 6 × 128 | Weights: 3 × 3 × 64 × 128<br>Bias: 1 × 1 × 128 |
| 15 | Activation | Function: ReLU | 123 × 6 × 128 | |
| 16 | Conv2D | Size: 3 × 3<br>No.: 128<br>Stride: 1 × 1 | 121 × 4 × 128 | Weights: 3 × 3 × 128 × 128<br>Bias: 1 × 1 × 128 |
| 17 | Activation | Function: ReLU | 121 × 4 × 128 | |
| 18 | Pooling | Type: Max Pooling<br>Size: 2 × 2<br>Stride: 2 × 2 | 61 × 2 × 128 | |
| 19 | Dropout | Ratio: 5% | 61 × 2 × 128 | |

(*continued*)

**Table 2.** (*continued*)

| Layer No. | Layer Type | Settings | Dimensions | Learnable Parameters |
|---|---|---|---|---|
| 20 | Fully Connected | No.: 2 048 | 1 × 1 × 2 048 | Weights: 2048 × 15616<br>Bias: 2048 × 1 |
| 21 | Activation | Function: ReLU | 1 × 1 × 2 048 | |
| 22 | Dropout | Ratio: 5% | 1 × 1 × 2 048 | |
| 23 | Fully Connected | No.: 512 | 1 × 1 × 512 | Weights: 512 × 2048<br>Bias: 512 × 1 |
| 24 | Activation | Function: ReLU | 1 × 1 × 512 | |
| 25 | Dropout | Ratio: 25% | 1 × 1 × 512 | |
| 26 | Fully Connected | No.: 5 | 1 × 1 × 5 | Weights: 5 × 512<br>Bias: 5 × 1 |
| 27 | Activation | Function: Softmax | 1 × 1 × 5 | |
| 28 | Output | | 1 × 1 × 5 | 28 |

participating in computations. This not only reduces computation time, but also transforms a single large-scale model into a collection of relatively smaller models, which effectively improves the model's ability to generalize.

## 4 Neural Network Training and Performance Analysis

### 4.1 Training Parameter Setting

Neural network training is the process of finding the weights between the convolutional kernels in the convolutional layers and the neurons in the fully connected layers, with the aim of minimizing the difference between the calculated output of the output layer and the true label given in the sample set. The selection and setting of the loss function and optimizer play a critical role in neural network training. First, the data samples are input into the neural network, then the current model performance is evaluated through the forward propagation process and the loss function. Next, the optimizer updates the weights of the learnable parameters in the neural network based on the size of the loss, using the backward propagation process.

The loss function used for training CBQNet is the cross-entropy loss function, which measures the distance between two probability distributions. Its expression is as follows:

$$C = -\frac{1}{n}\sum_{k=1}^{n}\left[y_k \ln a_k + (1 - y_k)\ln(1 - a_k)\right] \tag{3}$$

In this study, the optimizer used is Adadelta, an improved and extended version of the Adagrad algorithm. Compared with Adagrad, Adadelta no longer accumulates all past gradients, but adjusts the learning rate based on the moving window updated by the gradient, making it more robust. The main parameters for setting the Adadelta algorithm include a learning rate of 1.0, a decay rate of 0.95 for the moving average of gradient squares, a blur factor of $1 \times 10^{-6}$, and a learning rate decay value of 0 after each parameter update.

During the training process, 20% of the samples were randomly selected as the validation set, and the remaining 80% of the samples were used as training data. The total number of training epochs was set to 30, and 100 samples were fed into the neural network for each training iteration. The training environment was set up using Keras and TensorFlow. The workstation was equipped with an Intel Xeon E5-2673 v3 12C/24T 2.40 GHz processor and 64 G 2 400 MHz DDR4 ECC memory.

As each input image in the wellbore cementing quality evaluation sample set corresponds to two labels, representing the cementing quality of the first and second interfaces, respectively, two training processes are required during neural network training. The first training process uses the cementing quality of the first interface as the output, resulting in the CBQNet-1 neural network model for analyzing the quality of the first interface. The second training process uses the cementing quality of the second interface as the output, resulting in the CBQNet-2 neural network model for analyzing the quality of the second interface.

### 4.2 Performance Analysis

The accuracy and loss of CBQNet-1 during training are shown in Fig. 3 and Fig. 4. It can be seen from Fig. 3 that after the first training epoch, the model's training accuracy and validation accuracy were 78.5% and 84.2%, respectively, with a significant gap between them, indicating that the training was not sufficient. With the increase of training epochs, the training accuracy of the model showed a stable upward trend, with a fast-then-slow increase rate, and the training accuracy had exceeded 99% after 20 epochs. The upward trend of validation accuracy was consistent with that of training accuracy before the 12th epoch, and then validation accuracy showed some fluctuations without significant improvement. After 12 epochs of training, the training accuracy and validation accuracy of the model were 95.9% and 95.4%, respectively. Although further training could still improve the training accuracy, the validation accuracy no longer improved significantly, and the gap between the two began to increase, indicating that further training would lead the model to overfitting. From Fig. 4 we can see that the trend of the training loss and validation loss during training was basically the same as that of the accuracy, further indicating that the ideal effect could be achieved after 12 epochs of training.

The accuracy and loss of CBQNet-2 during training are shown in Fig. 5 and Fig. 6. It can be observed from Fig. 5 that after the completion of the first epoch, the model's training accuracy and validation accuracy were 40.3% and 46.7%, respectively, which were relatively low, indicating that a single round of training was insufficient for the neural network to fully grasp the rules between sample inputs and outputs. Similar to CBQNet-1, the model's training accuracy increased rapidly at first and then slowed down as the number of training epochs increased. After 20 epochs, the training accuracy

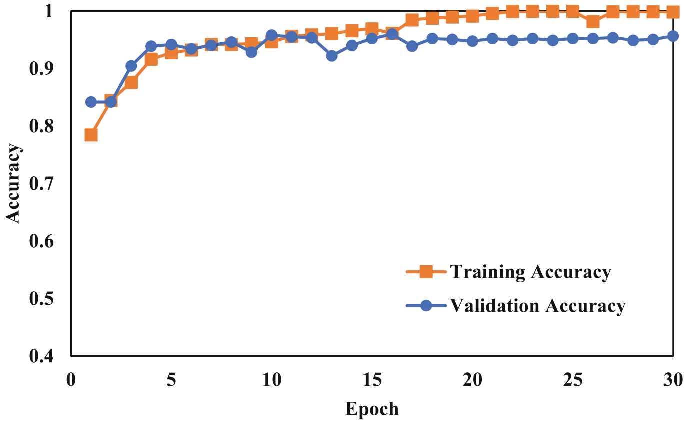


**Fig. 3.** The accuracy of CBQNet-1 during training.

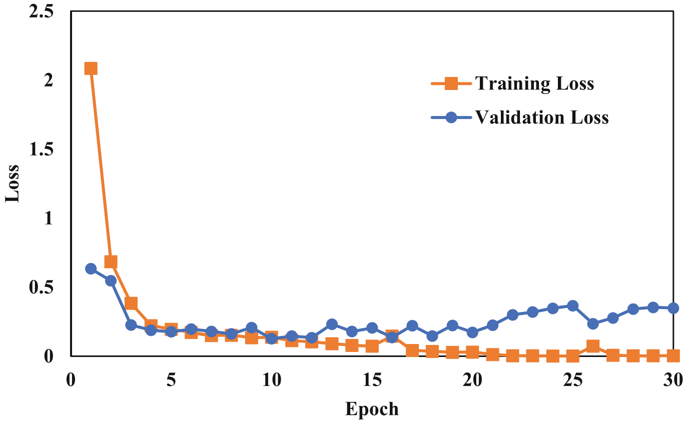


**Fig. 4.** The loss of CBQNet-1 during training.

exceeded 99%. The trend of the validation accuracy was consistent with that of the training accuracy before the 12th epoch. However, the validation accuracy showed a certain degree of fluctuation thereafter, with no significant improvement. After 12 epochs of training, the model's training accuracy and validation accuracy were 90.8% and 88.1%, respectively. From Fig. 6 we can see that the trend of the model's training loss and validation loss with respect to the number of training epochs was similar to the trend of the accuracy, which further indicates that the desired effect can be achieved after 12 epochs of training.

Overall, the accuracy of CBQNet-2 is lower than that of CBQNet-1, indicating that the analysis of the second interface bonding quality is more difficult than the analysis of the first interface bonding quality, which is consistent with the traditional understanding of manual analysis.

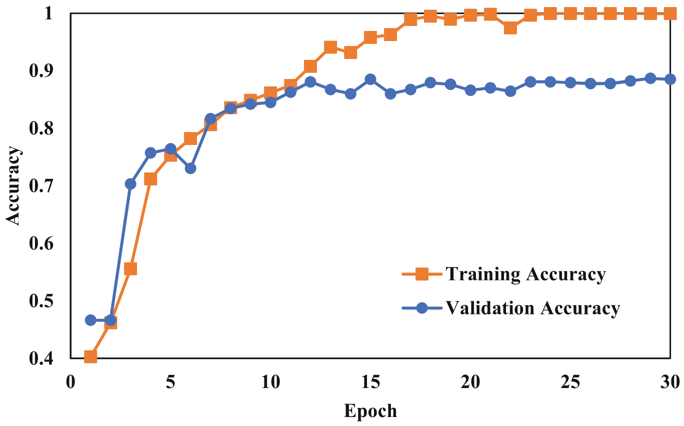


**Fig. 5.** The accuracy of CBQNet-2 during training.

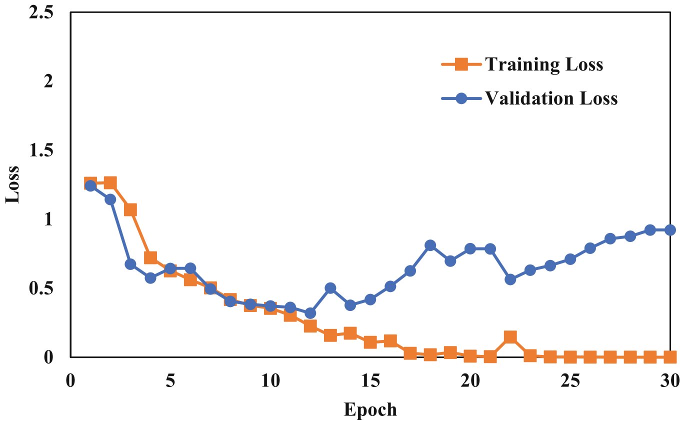


**Fig. 6.** The loss of CBQNet-2 during training.

The training time of CBQNet-1 and CBQNet-2 is shown in Fig. 7. From the figure, it can be observed that the training time of the two neural networks follows a similar trend. When the number of training epochs is small, the fluctuation in training time is stronger. However, with the increase in the number of training epochs, the training time of each epoch becomes more stable. The average training time per epoch is 188 s, indicating that the model's training efficiency is relatively high. If more samples are added in the future, it is possible to complete the training of a new model in a relatively short time.

Overall, the trained CBQNet-1 and CBQNet-2 can achieve high accuracy and automated analysis of the first and second interface cementing quality. They can save a lot of time spent on manual analysis, freeing petroleum engineers from simple and complicated work and allowing them to devote more energy to higher-level intelligent tasks such as operation management and anomaly handling.

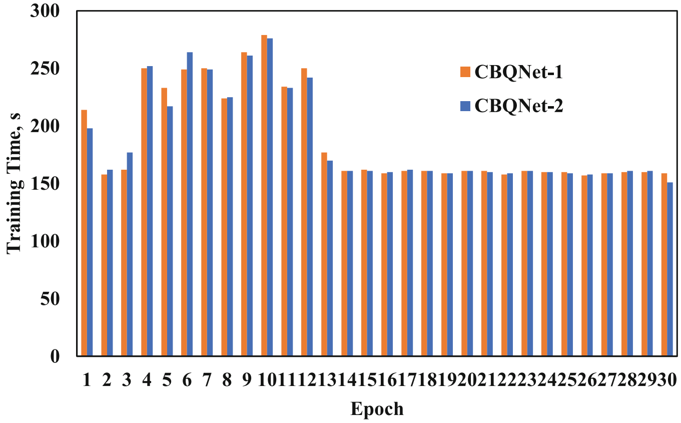


**Fig. 7.** The training time of CBQNet-1 and CBQNet-2.

## 5 Conclusion

A batch of historical cementing interpretation result images of oilfields were collected, and the images were standardized to establish a cement bond quality evaluation sample set. The sample set contains a total of 3351 samples, and each sample contains two labels of the cementing quality of the first interface and the second interface.

Combined with the characteristics of the cementing quality evaluation problem, the convolutional neural network was selected to carry out the personalized design of the network architecture, and a CBQNet with 28 layers and more than 32 million learnable parameters was constructed. After setting reasonable learning parameters, the CBQNet was trained with the ementing quality evaluation sample set, resulting in two models: CBQNet-1 for the cementing quality evaluation of the first interface and CBQNet-2 for the cementing quality evaluation of the second interface, with validation accuracy rates of 95.4% and 88.1%, respectively.

Future work will focus on expanding the cementing quality evaluation sample set, addressing the problem of uneven sample distribution, introducing more evaluation indicators, and further improving model accuracy.

**Acknowledgments.** The project is supported by National Natural Science Foundation of China (Number 52204027).

## References

1. Bigelow, E.L.: A practical approach to the interpretation of cement bond logs. J. Petrol. Technol. **37**(07), 1285–1294 (1985)
2. Jun, T., Zhang, C., Zhang, B., Fangfang, S.H.I.: Cement bond quality evaluation based on acoustic variable density logging. Petrol. Explor. Dev. **43**(3), 514–521 (2016)
3. Zuo, C., Qiao, W., Che, X., Yang, S.: Evaluation of azimuth cement bond quality based on the arcuate phased array acoustic receiver station. J. Petrol. Sci. Eng. **195**, 107902 (2020)

4. He, X., Chen, H., Wang, X.: Ultrasonic leaky flexural waves in multilayered media: cement bond detection for cased wellbores. Geophysics **79**(2), A7–A11 (2014)
5. Imrie, A.: The application of pattern recognition and machine learning to determine cement channeling & bond quality from azimuthal cement bond logs. In: SPWLA 62nd Annual Logging Symposium. OnePetro (2021)
6. Santos, L., Dahi Taleghani, A.: On quantitative assessment of effective cement bonding to guarantee wellbore integrity. J. Energy Resour. Technol. **144**(1) (2022)
7. Song, R.L., Liu, J.S., Lv, X.M., Yang, X.T., Wang, K.X., Sun, L.: Effects of tool eccentralization on cement-bond-log measurements: numerical and experimental results. Geophysics **78**(4), D181–D191 (2013)
8. Saini, P., Kumar, H., Gaur, T.: Cement bond evaluation using well logs: a case study in Raniganj Block Durgapur, West Bengal, India. J. Petrol. Explor. Prod. **11**, 1743–1749 (2021)
9. Nath, F., Kimanzi, R.J., Mokhtari, M., Salehi, S.: A novel method to investigate cement-casing bonding using digital image correlation. J. Petrol. Sci. Eng. **166**, 482–489 (2018)
10. Carletti, V., Greco, A., Percannella, G., Vento, M.: Age from faces in the deep learning revolution. IEEE Trans. Pattern Anal. Mach. Intell. **42**(9), 2113–2132 (2019)
11. Al-Naser, A., Al-Habib, M.: Adopting the fourth industrial revolution in oil and gas exploration. In: 81st EAGE Conference and Exhibition 2019, vol. 2019, no. 1, pp. 1–5. EAGE Publications BV (2019)
12. Suicmez, V.S.: What does the data revolution offer the oil industry? J. Petrol. Technol. **71**(03), 33 (2019)
13. Wang, H., Tao, G., Shang, X.: Understanding acoustic methods for cement bond logging. J. Acoust. Soc. Am. **139**(5), 2407–2416 (2016)
14. Gu, J., et al.: Recent advances in convolutional neural networks. Pattern Recogn. **77**, 354–377 (2018)
15. Sultana, F., Sufian, A. and Dutta, P.: Advancements in image classification using convolutional neural network. In: 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), pp. 122–129. IEEE (2018)

# Intelligent Prediction Technology for Production of Tight Oil Based on Data Analysis

Ning Li[1,2](✉), Xiang-hong Wu[1,2], Xin Li[1,2], Zhi-ping Wang[1,2], Yue-zhong Wang[1,2], Li-ao Zhao[1,2], Liang Ren[1,2], Hong-liang Wang[1,2], Hong-yu Tian[1], Shu-hang Ren[1,2], and Si-rui Jiang[3]

[1] Research Center of Artificial Intelligence, Research Institute of Petroleum Exploration and Development, PetroChina, Beijing, China
lining_riped@petrochina.com.cn

[2] Artificial Intelligence Technology R&D Center for Exploration and Development, CNPC, Beijing 100083, China

[3] Department of Asia-Pacific E&P, Research Institute of Petroleum Exploration and Development, PetroChina, Beijing, China

**Abstract.** China is rich in tight oil resources, with a wide distribution range and a large amount of resources, making it one of the key areas for strategic replacement of future oil reserves and production. In response to issues such as strong heterogeneity of terrestrial tight oil reservoirs, difficulty in drilling high-quality oil layers, large production differences, and unclear main control factors for production capacity, a detailed analysis of dynamic and static data of production wells was conducted to analyze production performance and decline patterns. Production wells were classified according to production characteristics, and development indicators at different stages were statistically analyzed based on actual production days. Using a combination of principal component analysis and Pearson correlation coefficient, based on multiple dynamic and static data such as geological factors, fracturing factors, and development factors, and analyzing the correlation between different single and combined factors and cumulative oil production at different stages, the main control factors for different production stages of tight oil were obtained. A production capacity prediction model for tight oil fracturing horizontal wells was established based on machine learning intelligent algorithms,

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20-22 September 2023.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 78–98, 2024.
https://doi.org/10.1007/978-981-97-0272-5_7

A production capacity evaluation and prediction technology for tight oil fracturing horizontal wells has been developed. By comparing with actual production data, the accuracy of the predicted results can meet production needs, providing a strong technical foundation for precise prediction and guidance of tight oil production in China.

**Keywords:** Tight oil · Production forecast · Data analysis · Analysis of main control factors · Intelligent algorithms

## 1 Introduction

Rich tight oil resources have been discovered in terrestrial sedimentary reservoirs of multiple basins in China, with a total resource volume exceeding 11 billion tons, making tight oil a major development replacement field and a new strategic growth point for crude oil production in China. Compared with North American marine tight oil, Chinese terrestrial tight oil has the characteristics of "multiple types, low porosity, low fluidity, and relatively poor oil properties". The geological conditions of terrestrial tight oil in China are complex, with multiple types and complex resource composition. The distribution of sand bodies is scattered, the vertical and horizontal continuity of reservoirs is poor, the reservoirs are dense and heterogeneous, and there are significant differences in single well drilling rates. The source reservoir relationship is mainly dominated by the intra source type, accounting for approximately 77.7%, the sub source type accounting for 18.2%, and the above source type accounting for 4.1%. The lithology is mainly composed of sandstone, accounting for about 69%, carbonate rock accounting for about 29.8%, and sedimentary volcanic rock accounting for about 1.3%. The pressure coefficient is mainly high pressure, 64.8% of which is $>1.2$, 29.3% of which is 0.8–1.2, and 22% of which is $<0.8$. The physical properties of crude oil are mainly low viscosity crude oil, with 41.2% having a viscosity of $<2$ mPa.s, 31.7% having a viscosity of 2–10 mPa.s, and 27.1% having a viscosity of $>10$ mPa.s.

Through the analysis of development effectiveness, domestic tight oil development currently faces two challenges in terms of production and efficiency: firstly, the large difference in single well production capacity, rapid decline, and low EUR of tight oil, which poses challenges to the effective utilization of tight oil. The second is the high cost and poor efficiency of using horizontal wells and volume fracturing for development. In the current context of low oil prices, how to reduce costs and improve development efficiency faces serious challenges. Through research, it has been found that the strong heterogeneity of the physical properties and oil-bearing properties of tight oil reservoirs is the fundamental reason for the significant productivity differences in horizontal wells. The significant difference in the effectiveness of tight oil fracturing is an important factor affecting production capacity. The production of tight oil in a single well depends on the production of each fracturing section, which is mainly controlled by the oil-bearing, physical properties, fluid properties, and fracturing effect of the reservoir; The organic matching of high-quality reservoir drilling rate and effective fracturing interval number is the main controlling factor for single well productivity. The strong heterogeneity of the reservoir is an important factor affecting the drilling rate. The low drilling rate and low saturation of movable fluids in Class I high-quality reservoirs are the fundamental reasons for the failure to achieve the expected production of horizontal wells.

In order to effectively predict the decline law of tight oil production, analytical and numerical calculation methods are currently mainly used. Among them, analytical calculation methods mainly include Arps decline curve method, typical decline curve chart method, relative permeability curve method, etc. Numerical calculation methods mainly refer to reservoir numerical simulation methods. However, each of these two methods has its advantages and disadvantages: the analytical calculation method has a fast calculation speed and can quickly provide a rough curve trend pattern. However, the decline pattern of tight oil is complex, and a single decline pattern formula is difficult to describe the overall decline process, and the calculation accuracy is not very accurate. However, the reservoir numerical simulation method can accurately calculate numerical solutions, but generally takes a long time and has high calculation costs.

With the gradual rise of artificial intelligence technology and the significant improvement of computer computing power, artificial intelligence prediction technology has emerged. From the perspective of big data analysis, this technology considers more influencing factors and is more comprehensive compared to traditional analytical methods. At the same time, compared to reservoir numerical simulation methods, it does not require global direct numerical simulation of the flow field values at each time step, greatly improving the calculation speed. Hamid Rahmanifard [1] made a detailed comparative analysis of the performance of ML algorithms and statistical methods, and then used two statistical methods (exponential smoothing and seasonal autoregressive comprehensive moving average) to make a comparative study of six kinds of modern ML networks, including multilayer perceptron (MLP), long short-term memory (LSTM), bidirectional LSTM (BiLSTM), convolutional neural network (CNN), long-term recursive convolutional network (LRCN) and gated recursive unit (GRU). In order to determine the relationship between static and dynamic data of some development units in the oilfield and the decline rate of oil production, Zhang Yan [2] used data-driven methods to identify the correlation between post fracturing production and production influencing factors by analyzing the geological properties and fracturing construction parameters of tight sandstone in Changqing Oilfield. Elastic networks, decision tree regression, support vector regression have been used to establish prediction models from reservoir properties and fracturing construction parameters to production. Liang Tao [3] established an initial cumulative oil production mixing model for Multi Fractured Horizontal Wells (MFHWs) that considers both geological and volumetric fracturing factors. Based on big data, a multi-level evaluation system has been established using Analytic Hierarchy Process. Calculate the weighting factor to reveal the key factors affecting the productivity of MFHWs. Using fuzzy logic method to calculate Euclidean distance and quantitatively predict the production of any horizontal well. Zainab Al Ali Hussain Al Ali [4] used two deep learning models, namely, Long short-term memory (LSTM) and N-BEATS, to predict the oil recovery data of two wells in Norway's Norne Oilfield. The use of pretrained N-BEATS models overcomes the shortcomings of LSTM models that previously required feature selection and rich training history, and the performance of N-BEATS meta learning methods is superior to LSTM models. The LSTM neural network model has been used multiple times to predict the trend of monthly oil production and water content in high water cut old oilfield blocks [5–11].

The eXtreme Gradient Boosting (XGBoost) algorithm is a scalable distributed gradient boosting decision tree (GBDT) machine learning library. XGBoost provides parallel tree enhancement function and is an advanced machine learning library for regression, classification, and ranking problems. XGBoost was initially initiated as a research project by Tianqi Chen as part of the Distributed (Deep) Machine Learning Community (DMLC) group. It is an optimized distributed gradient enhancement library designed for efficiency, flexibility, and portability. XGBoost is a tool for large-scale parallel boosting trees, which is more than 10 times faster than common toolkits. In terms of large-scale data in the industry, the distributed version of XGBoost has extensive portability, supporting running on various distributed environments such as Kubernetes, Hadoop, SGE, MPI, Dask, etc., making it a good solution to the problem of large-scale data in the industry.

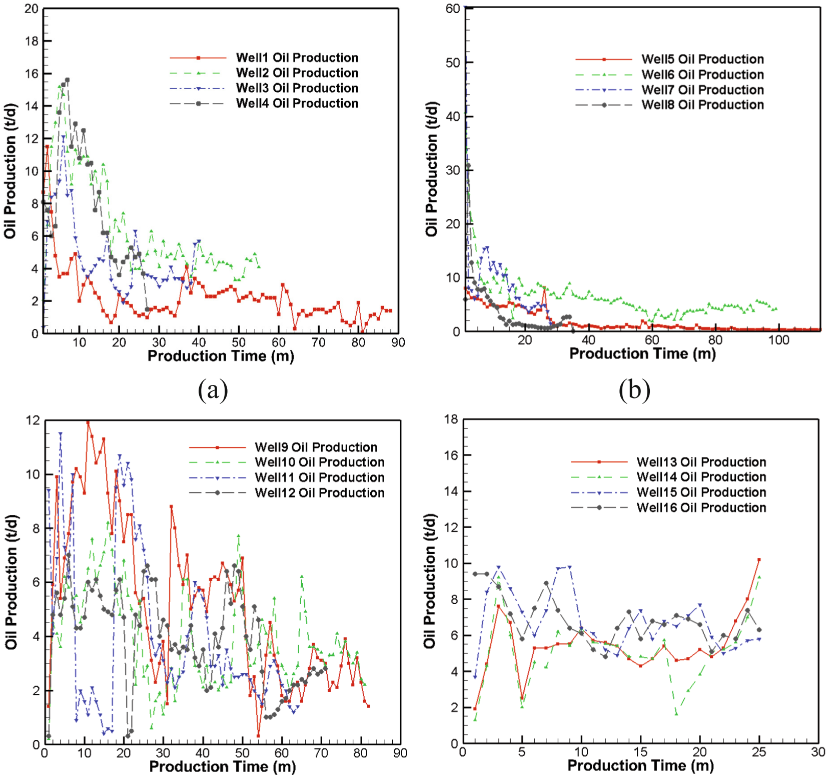
This paper adopts the XGBoost algorithm to establish a corresponding single well production decline prediction model based on the characteristics of tight oil reservoirs in China. Through practical application in a tight oil field in China, the superiority and correctness of this method in predicting single well production capacity have been confirmed, meeting the urgent needs of oilfield dynamic analysis, development planning, and decision-making.

## 2 Analysis of the Declining Law of Tight Oil Production

Although the overall changes in production characteristics of horizontal wells in each block are consistent, there are certain differences in the changes in daily liquid production, daily oil production, water content, production casing pressure, and other characteristics of each horizontal well based on the analysis of single well development performance data. Through literature research, it was found that most tight oil reservoirs are analyzed for production characteristics based on the variation of daily oil production with mining time. Therefore, this article will classify and analyze the production and mining characteristics of horizontal wells in the study area based on the variation of daily oil production. According to the curve characteristics of the daily oil production of a single well changing with mining time, the production and mining characteristics of horizontal wells can be divided into four categories:

### 2.1 Type 1: Rapid Increase in Initial Production and Short Stable Production Period

The overall performance is that the daily oil production capacity of horizontal wells continues to increase in the initial stage of production, and reaches the highest daily oil production level (10t–15t) within about 10 months. However, the stable production period is relatively short, and after 1 year of production, the daily oil production begins to decrease. After 2 and a half years, the daily oil production of a single well decreases to about 5t; The change in daily liquid production is similar to that of daily oil production; The water cut changes in the opposite direction and fluctuates within the range of 80% to 100%. The fracture network formed by horizontal well fracturing is the reason for high production in the initial stage of production, and the high production period is generally maintained between the second month and the sixth month, after which it enters the decreasing stage. The typical production curve is shown in Fig. 1(a).



(a)

(b)

(c)

(d)

**Fig. 1.** Curve of Daily Oil Production of a Single Well Changing with Production Time.

## 2.2 Type 2: High Initial Production and Rapid Decline in Later Stages

The overall performance is that the horizontal well has a high daily oil production capacity in the early stage of production, but the stable production period is extremely short. Generally, the daily oil production level starts to decrease within one month, and the daily oil production in the first three months drops to about 50% of the initial production, with a very fast decline rate. Generally, the daily oil production of the well drops to below 5t within 2–3 years of production. The changes in daily liquid production and production casing pressure of horizontal wells are similar to daily oil production. The typical production curve is shown in Fig. 1(b).

### 2.3 Type 3: The Fluctuation of Production is Large and Showing Multiple "Peaks"

The performance is that the daily oil production capacity of horizontal wells gradually increases in the initial stage of production, but the stable production period is short. During the production time, the daily oil production continuously fluctuates up and down. Overall, the daily oil production level is the strongest in the initial stage, and the daily production in the later stage shows a downward trend fluctuation. Generally, the daily production of wells decreases to below 5t after 4–5 years of production. The typical production curve is shown in Fig. 1(c).

### 2.4 Type 4: No Significant Fluctuations in Production and Maintaining Stable Production

The performance is that the daily oil production capacity of horizontal wells gradually increases in the initial stage of production, reaching its maximum in about 3 months, and the daily oil production is relatively stable throughout the entire production period, maintaining between 5–10t/d; The changes in water content and daily liquid production are similar to the daily oil production. The maximum daily oil production of this type of horizontal well is within the range of 5–10t/d, which is at a moderate level. At the same time, the production time is relatively short, mostly within two years. The daily production is still in a stable period, so there is no significant fluctuation and stable production has been maintained. The typical production curve is shown in Fig. 1(d).

## 3 Introduction of XGBoost Algorithm

XGBoost, as one of the Boosting algorithms, is a lifting tree model that integrates many tree models. By adding a regular term to the loss function, the complexity of the model is controlled to prevent overfitting. It can achieve parallel processing, which has greatly improved the speed compared to GBDT. XGBoost is essentially $k$ decision trees ($k$ is a positive integer), and the output of the regression tree is a real number (continuous variable). Boosting method is to combine multiple weak learners to give the final learning results, and take the output results of each weak learner as continuous values. The purpose of this is to accumulate the results of each weak learner, and better use the loss function to optimize the model.

Let $f^t(x_i)$ is the output result of the t-round weak learner, $\hat{y}_i^{(t)}$ it is the output result of the model, $y_i$ it is the actual output result, and the expression is as follows:

$$\hat{y}_i^{(t)} = \sum\nolimits_{k=1}^{t} f^k(x_i) = \hat{y}_i^{(t-1)} + f^t(x_i) \tag{1}$$

The objective function, that is, the loss function, builds the optimal model by minimizing the loss function. The loss function should add a regular term representing the complexity of the model, and the model corresponding to XGBoost contains multiple CART trees. Therefore, the objective function of the model is:

$$obj(\theta) = \sum\nolimits_{i}^{n} L\Big(y_i, \hat{y}_i^{(t)}\Big) + \sum\nolimits_{k=1}^{t} \Omega\Big[f^k(x_i)\Big] \tag{2}$$

The above formula is the regularization loss function. The first part on the right side of the equation is the training error of the model, and the second part is the regularization term. The regularization term here is the sum of the regularization terms of $k$ trees. The specific form is:

$$\Omega\left[f^k(x_i)\right] = \gamma T + \frac{1}{2}\lambda\|w\|^2 \tag{3}$$

where: $T$ is the number of leaf nodes, $\|w\|$ is the modulus of the leaf node vector, $\gamma$ it indicates the difficulty of node segmentation, indicates L2 regularization coefficient.

According to the expansion rule of the second derivative of the Taylor formula, the training error is further deduced and expanded to obtain:

$$obj(\theta)^{(t)} = \sum\nolimits_{j=1}^{T}\left[G_j w_j + \frac{1}{2}\left(H_j + \lambda\right)w_j^2\right] + \gamma T \tag{4}$$

where: $G_j$ represents the sum of the first derivative of all input samples mapped as leaf node $j$, $H_j$ represents the sum of second derivative of all input samples mapped to leaf node $j$.

In summary, we have introduced the main algorithms of XGBoost, which lays a theoretical foundation for subsequent prediction applications.

## 4 Workflow

For the prediction of well production in tight oil fields, first of all, data collection and preprocessing should be carried out, including the static and dynamic data of the reservoir, and the corresponding sample database should be established. Then, closely combining with the field data of the oilfield, and making full use of geological, engineering and development data, based on the production performance analysis and production decline law analysis in the study area, Identify the relevant influencing factors that affect the production capacity of horizontal wells for volume fracturing in tight oil reservoirs, calculate the partial correlation coefficient between the two factors, screen out independent influencing factors, and conduct single factor and multiple combination factor analysis from three aspects: geological parameters, engineering parameters, and development factors. Through Principal Component Analysis (PCA) & Pearson Correlation Coefficient Analysis method (PCCA) methods, comprehensively analyze multiple/single factors to screen out the main controlling factors for production capacity; Establish a prediction model based on XGBoost, which requires training and tuning the model to ultimately form the optimal XGBoost tight oil field well production prediction model. The specific process is shown in Fig. 2.

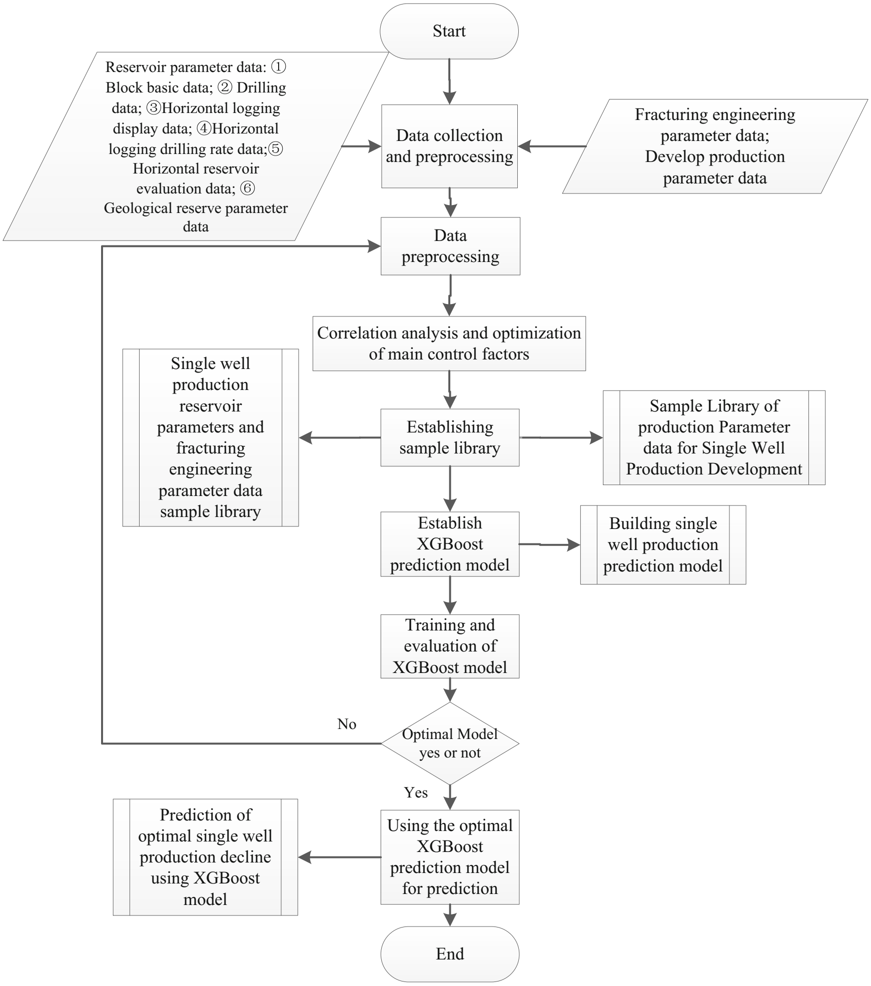

Start
Reservoir parameter data: ① Block basic data; ② Drilling data; ③Horizontal logging display data; ④Horizontal logging drilling rate data;⑤ Horizontal reservoir evaluation data; ⑥ Geological reserve parameter data
Data collection and preprocessing
Fracturing engineering parameter data; Develop production parameter data
Data preprocessing
Correlation analysis and optimization of main control factors
Single well production reservoir parameters and fracturing engineering parameter data sample library
Establishing sample library
Sample Library of production Parameter data for Single Well Production Development
Establish XGBoost prediction model
Building single well production prediction model
Training and evaluation of XGBoost model
No
Optimal Model yes or not
Yes
Prediction of optimal single well production decline using XGBoost model
Using the optimal XGBoost prediction model for prediction
End

**Fig. 2.** Technical workflow.

### 4.1 Data Collection and Preprocessing

**Collection and Organization of Data.** The production of a single well in a tight oil field is influenced by various factors, mainly including reservoir parameter data, fracturing engineering parameter data, and development and production parameter data. In terms of reservoir parameter data, it also includes block basic data, drilling data, horizontal section logging display data, horizontal section drilling rate data, horizontal section reservoir evaluation data, geological reserve parameter data, etc. The specific relevant parameters are shown in Table 1.

**Table 1.** Collected dynamic and static parameters.

| Data classification | | Related parameters |
|---|---|---|
| Reservoir parameter data | Block basic data | a) Block, well number, horizon, sublayer, reference well, interpretation layer, effective thickness of each sublayer, designed production capacity, designed well depth, and designed horizontal section length |
| | Drilling data | b) First drilling time, completion time, year of production, completion method, cycle, well depth, artificial bottom hole, oblique depth of point A during actual drilling, vertical depth of point A during actual drilling, length of horizontal section, length of horizontal section used |
| | Horizontal logging display data | c) The length of sandstone encountered in horizontal section logging, the length of oil layer encountered in horizontal section logging, the drilling rate of sandstone encountered in horizontal section logging, the drilling rate of oil layer encountered in horizontal section logging, the oil immersion length of horizontal section logging, the length of oil spot in horizontal section logging, the length of oil stains in horizontal section logging, the fluorescence length of horizontal section logging, and the total length of horizontal section logging |

(*continued*)

**Table 1.** (*continued*)

| Data classification | | Related parameters |
|---|---|---|
| | Horizontal logging drilling rate data | d) The length of sandstone encountered during horizontal logging, the length of oil layer encountered during horizontal logging, the drilling rate of sand-stone encountered during horizontal logging, and the drilling rate of oil layer encountered during horizontal logging |
| | Horizontal reservoir evaluation data | e) Horizontal Section I Reservoir Length, Horizontal Section II Reservoir Length, Horizontal Section III Reservoir Length, Horizontal Section IV Reservoir Length, Horizontal Well Classification Evaluation Category, Horizontal Well Classification Evaluation Index |
| | Geological reserve parameter data | f) Production thickness, fracture length, porosity, average saturation, density, volume coefficient, controlled reserves, production reserves |
| Fracturing engineering parameter data | | g) Fracturing completion structure, number of fracturing segments, number of fracturing clusters, average interval spacing, average cluster spacing, total fracturing fluid volume, total fracturing sand volume, fluid intensity, sand intensity, single stage fluid volume, single stage sand volume, fracturing completion time, soaking time after fracturing, single cluster fluid volume, single cluster sand volume |

(*continued*)

**Table 1.** (*continued*)

| Data classification | Related parameters |
|---|---|
| Develop production parameter data | h) Block, well number, well pattern type, production time, oil production method, pump diameter, oil pressure, casing pressure, dynamic liquid level, production days, cumulative production days, monthly oil production, monthly water production, verification of monthly oil production, verification of monthly water production, verification of cumulative oil production, verification of cumulative liquid production, number of wells opened, daily liquid production capacity, daily oil production capacity, water content, verification of daily liquid production capacity, verification of daily oil production capacity, depth of middle oil layer, flow pressure, storage and production coefficient Return rate, deficit, recovery degree, upward pumping time, and number of months of self-production |

In addition to the single factor mentioned above, in order to highlight the impact of different factors and have a greater correlation with production capacity, the following multiple factors have been added according to the needs of the research problem, including:

Among them, the effective length of the horizontal well $L_{eh}$ is

$$L_{eh} = a_1 L_{oi} + a_2 L_{osp} + a_3 L_{ost} + a_4 L_f \tag{5}$$

And the oil-bearing $S_{ob}$ is

$$S_{ob} = \frac{L_{eh}}{L_{oi} + L_{osp} + L_{ost} + L_f} \tag{6}$$

where: $L_{oi}$ stands for the length of oil immersion, m; $L_{osp}$ stands for the length of oil spot, m; $L_{ost}$ stands for the length of oil stains, m; $L_f$ stands for the length of fluorescence, m; $a_i$, $i = 1, 2, 3, 4$ stands for the weight.

**Table 2.** Added multiple factor parameters.

| Data classification | Related parameters |
|---|---|
| Reservoir parameter multiple data | a) Oil-bearing<br>b) Effective length of horizontal well<br>c) Permeability × Thickness used<br>d) Permeability × Thickness used × Effective length of horizontal well<br>e) Utilized reserves × Permeability<br>f) Utilized reserves × Permeability/Viscosity<br>g) Oil-bearing × Production reserves × Permeability/Viscosity<br>h) Reservoir quality × Oil-bearing × Produced reserves × Permeability/Viscosity |
| Multiple data of fracturing engineering | a) Liquid strength × Sand strength<br>b) Liquid strength × Sand strength × Number of segments<br>c) Liquid strength × Sand strength × Number of segments × Effective length of horizontal well<br>d) The amount of liquid added in single stage × Effective length of horizontal well<br>e) The amount of sand added in single stage × Effective length of horizontal well |

Note: Liquid strength equals to the amount of total liquid/Utilized reserves. Sand strength equals to the amount of total sand/Utilized reserves

**Data Preprocessing.** For different types of data in tight oil well areas, data cleaning is carried out based on their data volume, data type, data quality, etc., eliminating duplicate well information, completing missing data, data integration, data transformation, and other processes, and corresponding preprocessing is carried out for each data item.

(1) Correction of flowback period data: After fracturing construction, the production during the flowback period is very low, which is not a normal industrial oil flow. Therefore, it is necessary to remove the time period of the flowback period and the oil production below a certain amount. The specific quantitative values vary from different oilfields;
(2) Reorganize production data based on differences: Due to the cleaning of flowback period data and time periods, it is necessary to recalculate the cumulative oil production, cumulative liquid production, and water content for different production time periods;
(3) Removal of abnormal well data: Based on expert experience and data analysis, identify wells with abnormally high or low production by drawing charts, and eliminate them according to specific circumstances;
(4) Pre processing of specific data tables:
    (a) Reservoir static data: porosity, formation pressure, and other data, with fixed values for each block. Based on the collected data, these types of data are supplemented in the data table.

(b) Developing dynamic data: Dynamic data such as extraction degree and dynamic liquid level vary, varies with production time, and need to be recalculated and organized based on expert experience and specific formulas.
(c) The combination of dynamic and static data: Through difference calculation, the production dynamic data has been reorganized and calculated. Merge the newly generated development dynamic data into a static data table according to the well name.

### 4.2 Data Correlation Analysis

After sorting out the influencing factors of production capacity and preprocessing the data, it is necessary to conduct correlation analysis between the influencing factors and production capacity, and screen out the main controlling factors of production capacity. This article adopts a combination of principal component analysis (PCA) and Pearson correlation coefficient, the method of combining PCA and Pearson is adopted.

**Principal Component Analysis (PCA).** The principal component analysis method is to transform multiple existing indicators into a few well representative comprehensive indicators, which can reflect most of the information of the original indicators and maintain independence between each indicator to avoid overlapping information. Principal component analysis mainly plays a role in reducing dimensionality and simplifying data structures.

(a) Standardize indicator data, collect p-dimensional random vectors $X$, n samples,

$$X_i = \left\{X_{i1}, X_{i2}, \ldots, X_{ip}\right\}^T, (i = 1, 2, \ldots, n) \tag{7}$$

Construct a sample matrix and perform standardized transformation on the sample matrix;

$$Z_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j}, (i = 1, 2, \ldots, n; j = 1, 2, \ldots, p) \tag{8}$$

(b) Calculate correlation coefficient matrix based on standardized matrix;

$$R = \left[r_{ij}\right]_p xp = \frac{Z^T Z}{n - 1} \tag{9}$$

(c) Solve the characteristic equation of the sample correlation matrix $R$, obtain $p$ characteristic roots, and determine the principal components;

$$U_{ij} = z_i^T b_j^o, (j = 1, 2, \ldots, m) \tag{10}$$

(d) Convert the standardized indicator variables into main components;
(e) Perform a comprehensive evaluation of $m$ principal components, sum them with weights, and obtain the final evaluation value. The weight is the variance contribution rate of each principal component.

**Pearson Correlation Coefficient Analysis Method (PCCA).** Pearson correlation coefficient analysis is a method used to measure the degree of correlation between two variables $X$ and $Y$, with values between $-1$ and $+1$. Defined as the quotient of covariance and standard deviation between two variables.

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}} \tag{11}$$

By using the above method, the correlation coefficients between the factors in Tables 1 and 2 and the cumulative oil production at different stages were obtained, as shown in Table 3.

**Table 3.** Correlation analysis results of various factors and oil production.

| Influence factor | Correlation coefficient (Weight) |
|---|---|
| Effective length of horizontal section | 0.304 |
| Oil-bearing | 0.212 |
| Thickness used | 0.152 |
| Controlled reserves | 0.152 |
| Permeability × Thickness used | 0.05 |
| Controlled reserves × Permeability/Viscosity | 0.05 |
| Permeability × Thickness used /Viscosity | 0.015 |
| Number of fracturing segments | 0.222 |
| The amount of sand added in single cluster | 0.14 |
| Liquid strength | 0.098 |
| The amount of liquid added in single stage | 0.091 |
| Storage and production correlation coefficient | 0.08 |
| The amount of sand added in single cluster | 0.079 |
| Post-pressure soaking time | 0.07 |
| Total amount of sand added | 0.05 |
| Production pressure difference | 0.048 |
| Return rate before oil breakthrough | 0.035 |
| Sanding strength | 0.023 |
| Liquid strength × Sand strength | 0.017 |
| Number of fracturing clusters | 0.017 |

(*continued*)

**Table 3.** (*continued*)

| Influence factor | Correlation coefficient (Weight) |
|---|---|
| Bottom hole pressure | 0.012 |
| The amount of sand added in single stage | 0.01 |
| Total amount of liquid added | 0.005 |
| Water displacement before oil exposure | 0.002 |

### 4.3 Selecting Main Controlling Factors for Tight Oil Production

Through the above data analysis and combined with expert experience, the following parameters were ultimately selected as the main control factors (Table 4):

**Table 4.** Results of main control factors for tight oil production.

| Classification | Main control factors |
|---|---|
| Main control factors for geology | a) Effective length of horizontal section |
| | b) Oil-bearing |
| | c) Thickness used |
| | d) Controlled reserves × Permeability/Viscosity |
| Main control factors for fracturing and development | e) Number of fracturing segments |
| | f) The amount of sand added in single cluster |
| | g) Liquid strength |
| | h) The amount of liquid added in single stage |
| | i) The amount of liquid added in single cluster |
| | j) Total amount of sand added |
| | k) Production pressure difference |

### 4.4 Constructing a Typical Well Production Sample Library

Combining professional knowledge and expert experience, based on correlation analysis results and combined with cumulative production data from different stages, a sample library reflecting the changes in single well production was established. Through the sample library, expert experience was reflected.

### 4.5 Establishing a Multi Parameter Prediction Model for Well Production

**Model Construction.** Due to the fact that this article only involves three blocks of an oil field with a small sample size, it belongs to the small sample problem. Therefore, in the design of the prediction model, the concept of cyclic input is considered, which is to establish production prediction models according to different stages. When predicting the current stage of production, the cumulative output value of the previous production stage is input, as follows (Table 5):

**Table 5.** Prediction model input and output values.

| No. | Input values | Output value |
|---|---|---|
| 1 | Main control factors | $Q_1$ |
| 2 | Main control factors, $Q_1$ | $Q_3$ |
| 3 | Main control factors, $Q_1$, $Q_3$ | $Q_6$ |
| 4 | Main control factors, $Q_1$, $Q_3$, $Q_6$ | $Q_9$ |
| 5 | Main control factors, $Q_3$, $Q_6$, $Q_9$ | $Q_{12}$ |
| 6 | Main control factors, $Q_6$, $Q_9$, $Q_{12}$ | $Q_{18}$ |
| 7 | Main control factors, $Q_9$, $Q_{12}$, $Q_{18}$ | $Q_{24}$ |
| 8 | Main control factors, $Q_{12}$, $Q_{18}$, $Q_{24}$ | $Q_{36}$ |
| 9 | Main control factors, $Q_{18}$, $Q_{24}$, $Q_{36}$ | $Q_{48}$ |

Note: $Q_i$ stands for the accumulated oil production until the $i^{th}$ month.

As shown in the above table, this article adopts the concept of "equal dimensional replenishment", which refers to the dimension of input data. Except for the initial three stages as the initiation stage, all other stages use fixed four dimensional data input, always using the latest stage production data as the input of the model, and establishing a mapping relationship with the accumulated oil production in the next stage.

**Model Training.** Configure algorithm parameters and conduct model training.

(1) Max_depth: The maximum depth of each tree. When establishing each tree, achieving the expected accuracy or maximum depth will proceed to the next tree model construction. The default value is 6.
(2) Learning rate: learning rate is one of the most important hyperparameter. After each new tree model is established, the prediction results of the new model are given based on the previous prediction results and the interaction between the leaf output and the learning rate calculated this time. For different problems, the ideal learning rate will fluctuate between 0.05 and 0.3.
(3) Booster model: There are two models to choose: gbtree and gblinear. Gbtree uses a tree based model for lifting calculations, while gblinear uses a linear model for lifting calculations. The default is gbtree.
(4) Gamma: The minimum "loss reduction" required for further splitting at leaf nodes, with a default of 0.

(5) Min_child_weight: It can be understood as the minimum number of samples for leaf nodes, with a default of 1.
(6) Subsample: The sampling ratio of the training set. Before fitting a tree, this sampling step will be performed, with a value range of (0, 1]. The default is 1.
(7) Colsample_bytree: Before fitting a tree each time, determine how many features to use, with a value range of [0, 1], and the default value is 1.
(8) Reg_alpha: Tuning of regularization parameters. The alpha parameter can reduce the complexity of the model, thereby improving its performance.
(9) Reg_lambda: Tuning regularization parameters. Lambda parameters can reduce the complexity of the model and improve its performance.
(10) Random_State: Random seed, 0 by default.

**Model Evaluation.** Based on parameters such as the error and root mean square error between the predicted and actual data of the model, model optimization is carried out to provide the optimal model for predicting single well oil production. The calculation method for model accuracy is:

(1) Calculate the data of individual well oil production over time for each well sample in the test set;
(2) Calculate the average absolute percentage error between all predicted data points and actual data points, which is the model prediction accuracy.

During the calculation process, the following error calculations were used [12–18]. Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE):

$$MAE = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left|y'_i - y_i\right| \tag{10}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y'_i - y_i\right|}{y_i} \tag{11}$$

Coefficient Determination ($R^2$):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - y'_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}_i\right)^2} \tag{12}$$

### 4.6 Using Optimal Intelligent Models for Indicator Prediction

In response to the problem of predicting single well oil production, the optimized and trained XGBoost prediction model for single well oil production was used to carry out prediction work, obtaining future trends of single well oil production that can be used to guide actual production and conform to production laws. This plays a positive guiding role in production operation scheduling and adjustment of work systems.

# 5 Calculation Results and Analysis

## 5.1 Overview of FY Oilfield Work Area

The FY oil layer is the earliest discovered, most abundant, and widely distributed oil layer in the southern part of the SL Basin. The FY oil layer is distributed in the CL depression, HG terrace, and western region of the FX uplift zone in the central depression area. FY oilfield includes three blocks: Q block, R1 block, and R2 block.

## 5.2 Establishing a Multi Parameter Intelligent Prediction Model for Single Well Indicators

**Model Construction.** The XGBoost model for predicting single well oil production in tight oil fields was constructed using the XGBoost model introduced in the previous section. Establish models for different production stages.

**Model Training.** According to the basic content of the model training parameters mentioned in the previous section, parameter tuning tests were conducted with the accuracy of the test set as the evaluation label. There are a total of 84 wells in the sample set, with a ratio of 8:2 for training + validation sets, and testing set. This means that there are a total of 67 wells in the training + validation set, and 17 wells in the testing set. Compared through testing, max_depth is 15, learning rate is 0.1, boost model is gbtree, gamma is 0, min_child_weight is 1, subsample is 1, colsample_bytree is 1, reg_alpha is 0, reg_lambda is 1, random_state is 0.

**Model Evaluation.** The prediction model is constructed based on different sample types, and the final 12 to 48 months prediction model $R^2$ has an average accuracy of 86%, an average MAPE value of 13%, and an average MAE value of 351t.

## 5.3 Model Prediction Results and Analysis Discussion

Based on the optimal oil production prediction model in this article, relevant prediction work was carried out for 17 wells in three blocks of FY Oilfield. The comparison between the predicted results and actual production data of four wells is listed below, as shown in Figs. 3 and 4.

Figure 3 shows the predicted results of cumulative oil production at different production stages of wells WQ1 and WQ2. It can be seen that at the beginning of production, the predicted results are in good agreement with the actual production curve. During the production period of 20 to 48 months, the predicted values were slightly higher than the actual production data. The predicted value of WQ2 well in the mid-term production stage is slightly lower than the actual production data, and then the predicted value and production value continue to increase by the same magnitude.

Figure 4 shows the results of cumulative oil production predictions for WQ3 and WQ4 wells at different production stages. It can be seen that the predicted value of WQ3 well is generally lower than the actual production value, but the difference is relatively small. However, in the early stage of production, the predicted value of WQ4 well

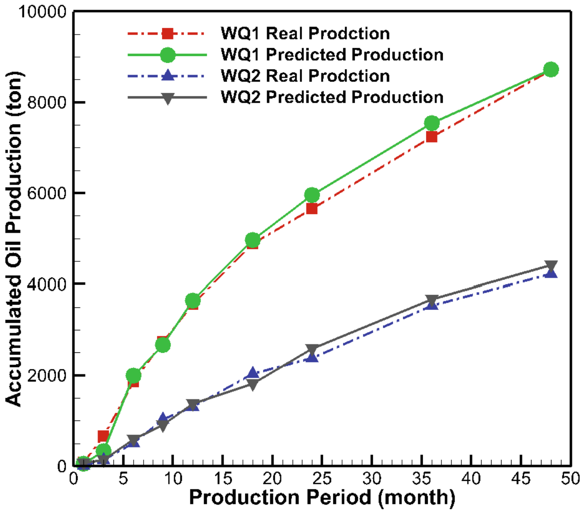


**Fig. 3.** Comparison between the predicted and real cumulative oil production with wells WQ1 and WQ2.

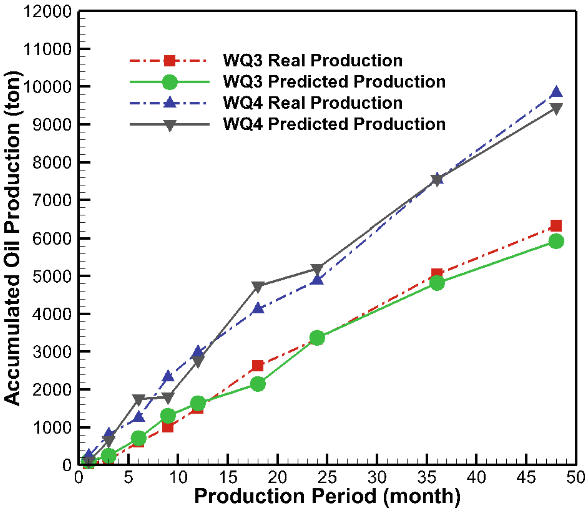


**Fig. 4.** Comparison between the predicted and real cumulative oil production with wells WQ3 and WQ4.

increases alternately with the actual production data, and remains basically consistent after 25 months of production.

From the comparison between the predicted results in Figs. 3 and 4 and the actual production curve trend, as well as the model error evaluation results, it can be seen that the prediction accuracy of the model in this paper is relatively high in predicting the cumulative oil production over 48 months. This indicates that the prediction model established through a series of methods and techniques introduced in this article is more effective in predicting the cumulative production of a single well, thus achieving multi-dimensional tight oil single well production prediction, This provides a strong technical foundation for precise prediction and reasonable optimization of tight oil production in China.

## 6 Conclusion

Based on the XGBoost model, a typical tight oil well production sample library was constructed through data collection, organization, and preprocessing. Correlation analysis of influencing factors was conducted, and a multi-parameter intelligent prediction model for single well oil production indicators was established. The development indicators were predicted, and the conclusion is as follows:

(1) Established a complete and effective method for predicting development indicators of tight oil fields based on XGBoost model;
(2) The XGBoost cumulative oil production prediction model established is suitable for predicting the trend of cumulative oil production in tight oil fields, and the established model has a high accuracy in predicting single well production;
(3) The methods and techniques introduced in this article are not only limited to tight oil fields, but can also be applied to the production prediction of unconventional oil and gas fields.

In summary, the artificial intelligence model established in this article has achieved multi-dimensional prediction of single well tight oil production, improved the dynamic management level of oil well production, improved the accuracy of single well measure decision-making, and improved the ultimate oil recovery rate and production efficiency of the oilfield. This provides a strong technical foundation for precise prediction of tight oil production and reasonable optimization of production allocation in China.

**Acknowledgments.** The work is supported by the following projects: ① Jilin Oilfield Branch project “Research on Capacity Evaluation and Parameter Optimization of Typical Tight Oil Blocks” (Number: JS2022-W-13-JZ-37-47); ② CNPC RIPED project “Research on integrated mathematical model of fracture network and software module development” (Number: YGJ2019-07-04). ③ CNPC major scientific and technological project of upstream key core technology “Volume fracturing optimization design software” sub-project “EDFM-based Component reservoir simulation software” (Number: 2020B-4118). ④ CNPC RIPED project “Joining 5 industrial alliance organizations such as the University of Calgary for exchange and cooperation research” sub-project “Joining the MMRD Consortium at Colorado School of Mines” (Number: 2021DQ0105-03). ⑤ CNPC Project “Research on Intelligent Interpretation Technology and Knowledge Graph of Oil and Gas Exploration” (Number: 2021DJ7003). And in the process of completing this paper, the author has received technical support and help from researchers, including Li Chen, Zhong-cheng Li, Hong-yu Zhao, Chang-chun Dong, Qi Liu, Yi-meng Wang of Jilin Oilfield Exploration and Development Research Institute, and experts, incluidng Bin Bai, Xiu-Lin Hou, Shu-jian Liu of Research Institute of Petroleum Exploration & Development PetroChina, Feng-peng Lai, Lin-lin Zhang of China University of Geosciences, and the author would like to express the sincere gratitude.

## References

1. Rahmanifard, H., Gates, I., Asl, A.S.: Comparison of machine learning and statistical predictive models for production time series forecasting in tight oil reservoirs. In: SPE/AAPG/SEG Unconventional Resources Technology Conference, Houston, Texas, USA (2022). https://doi.org/10.15530/urtec-2022-3703284
2. Zhang, Y., Zheng, Y., Sun, S., et al.: Data driven production prediction of tight sandstone after compression in Changqing Oilfield. Energy Environ. Protect. **43**(10), 96–101127 (2021)
3. Tao, L., et al.: A new productivity prediction hybrid model for multi fractured horizontal wells in tight oil reservoirs. In: SPE/IATMI Asia Pacific Oil&Gas Conference and Exhibition, Virtual (2021). https://doi.org/10.2118/205620-MS
4. Al Ali Hussain Al Ali, Z., Horne, R.: Meta learning using deep N-BEATS model for production forecasting with limited history. In: Gas&Oil Technology Showcase and Conference held in Dubai, UAE (2023)
5. Understanding LSTM Networks. https://web.stanford.edu/class/cs379c/archive/2018/
6. Class_Messages_Listing/content/Important_Neural_Network_Technology_Tutorials/Olah /LSTM Neural Network Tutorial-15.pdf
7. Wang, Y., Wang, C., Zhang, H., et al.: Automatic ship detection based on RetinaNet using multi resolution Gaofen-3 image. Remote Sens. **11**(5), 531 (2019)
8. Chen, L., Wang, Z., Wang, G.: Application of LSTM network in short-term power load forecasting under deep learning framework. Power Inf. Commun. Technol. **15**(5), 8–11 (2017)
9. Li, N., Gong, R., Liu, Z., Mi, L., Liu, L.: Application of artificial intelligence technology in single well production and water cut prediction. In: Lin, J. (ed.) IFEDC 2021. Springer Series in Geomechanics and Geoengineering, pp. 512–528. Springer, Singapore (2021). https://doi.org/10.1007/978-981-19-2149-0_47
10. Ma, Q., Guo, J., Li, N.: Load forecasting methods for urban gas pipeline networks. J. Anshan Univ. Sci. Technol. **27**(2), 101–105 (2004)
11. Li, N.: Research on load forecasting of urban gas pipeline networks. Master's thesis, Liaoning University of Science and Technology (2004)
12. Ojedapo, B., Ikiensikama, S., Wachikwu, V.U.: Elechi petroleum production forecasting using machine learning algorithms. In: SPE Nigeria Annual International Conference and Exhibition held in Lagos, Nigeria (2022). https://doi.org/10.2118/212018-MS
13. Gong, R., Li, X., Li, N., et al.: Artificial Intelligence for Oil and Gas, pp. 9–10. Petroleum Industry Press (2021)
14. Li, N., Gong, R., Li, X., Li, W., Wu, B., Ren, S.: Factor analysis of affecting the accuracy for intelligent picking of seismic first arrivals with deep learning model. In: Lin, J. (ed.) IFEDC 2022. Springer Series in Geomechanics and Geoengineering, pp. 7042–7062. Springer, Singapore (2023). https://doi.org/10.1007/978-981-99-1964-2_598
15. Li, N., Li, L., Wu, S., Wu, Y.: Numerical simulation of the effect of nanocon-finement on hydrocarbon phase behavior in nanometer scale pores. In: Lin, J. (ed.) IFEDC 2019. Springer Series in Geomechanics and Geoengineering, pp. 162–174. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-2485-1_18
16. Li, N., Ran, Q.Q., Li, J.F., Yuan, J.R., Wang, C., Wu, Y.S.: A multiple-continuum model for simulation of gas production from shale gas reservoirs. SPE165991 (2013)
17. Li, N., Yan, L.: Direct numerical simulation of a mixed-media model for efficient developing shale gas reservoirs. In: Lin, J. (ed.) IFEDC 2020. Springer Series in Springer Series in Geomechanics and Geoengineering, pp. 1993–2009. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-0761-5_189
18. Li, N., Yan, L., Li, L., et al.: Numerical simulation of triple media percolation mechanism of shale gas reservoir. In: 10th National Symposium on Efficient Development Technology of Natural Gas Reservoir, pp. 342–349 (2019)

# Development Index Prediction Through Big Data Analysis for QX Ultra-Deep Permian Marine Carbonate Gas Reservoir in Sichuan Basin, China

Xiaohua Liu[1](✉), Xuliang Liu[2], Zhenhua Guo[1], Jichun Zhou[3], and Daolun Li[2]

[1] PetroChina Research Institute of Petroleum Exploration and Development, Xueyuan Road, No. 20, Haidian District, Beijing 100083, China
lxh69@petrochina.com.cn

[2] Hefei University of Technology, Hefei, Anhui, China

[3] Northwestern Sichuan Gas District, PetroChina Southwest Oil & Gas Company, Jiangyou, Sichuan, China

**Abstract.** Uncertainties in the characterization of new-found, ultra-deep, thin and low porosity Permian gas reservoir reduce feasibility for development index (DI) prediction through reservoir simulation. DI prediction with big data analysis approach are studied. Geology and production data from 30 mature gas fields are reviewed and 13 parameters are selected to represent geological features, deliverability and DI of individual reservoir. Based on the BP neural network algorithm, proxy models are established to correlate DI with geology and deliverability data, and the bagging method is used to effectively improve the experimental accuracy and stability while avoiding over-fitting phenomenon in the case of limited sample data. The coefficient of determination coefficient ($R^2$) are selected to evaluate the prediction effect of DI. The mean value of the prediction results of the model with higher $R^2$ value in 2000 numerical experiments was selected as the final prediction result. With the established proxy model, DI for QX reservoir in Permian formation are predicted and the influence of heterogeneity are also evaluated.

**Keywords:** Development index prediction · BP neural network · big data analysis · ultra-deep gas reservoir · Permian marine carbonate

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 99–112, 2024.
https://doi.org/10.1007/978-981-97-0272-5_8

## 1 Introduction

The Ultra-deep (>7000 m) Permian marine carbonate formation is a prospective conventional gas exploration and development domain in Sichuan Basin, China, and in recent years, significant breakthrough have been made in some appraisal wells with testing gas rates above 1.0 MMm$^3$/D from Middle Permian QX gas reservoir in Sichuan Basin, China (hereinafter referred to as QX reservoir). To meet market needs for clean energy, the full development of this reservoir is put on agenda.

The QX reservoir is structurally located at the Longmen Mountain buried thrust front zone, and contains many NE-SW trending faulted anticline and faulted nosing structures, and currently, 6 NE-SW trending tectonic high belts have been defined through seismic and a rough structure map is given in Fig. 1. Due to limited drillings and complex structure, the extension of faults, communication between each belt and gas and water distribution in QX reservoir are still uncertain, which reduce feasibility for a full reservoir modeling and simulation.

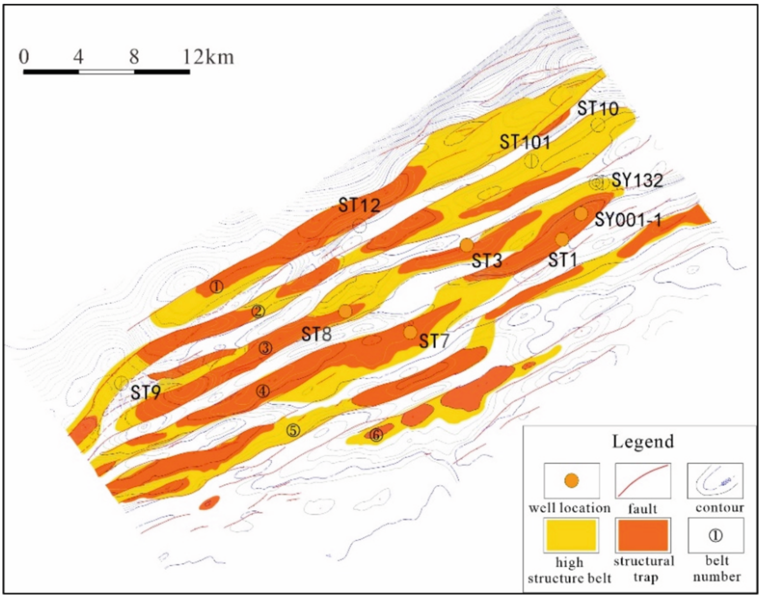


**Fig. 1.** Top structure map of Permian QX reservoir, SYS Block, Sichuan Basin

Current drilling, geology and geophysics studies show that QX reservoir is featured with ultra-deep buried depth (7200–7800 m), high pressure (>93 MPa), low porosity (3.9% in average) and thin layer (average pay zone thickness 20 m). Due to uneven development of natural fractures and vugs, heterogeneity exists in this reservoir with permeability ranging among 0.01–10 mD. The uncertainty and heterogeneity in this ultra-deep, thin layered and low porosity reservoir pose risks in cost-effective development, to lower risks in initiating exploitation activities, proper Development Index (DI) for guiding the commercial and steady development of the new findings are the key concerns of management.

Usually, in Field Development Plan (FDP), full reservoir modeling and simulation will be performed to predict DI which consists of a series of parameters including Field Annual Production Rate (FAPR), Field Plateau Period (FPP) at certain FAPR, Well Spacing Density (WSD), Well Average Daily Production (WADP) during FPP, and field Ultimate Recovery Factor (URF). And these key index will direct operators to make drilling plan and development policies. But this common approach is less reliable in QX reservoir due to uncertainties in reservoir characterization. Recently, big data analysis technique (Safavian et al., 1991; Quinlan, 1986; Rao et al., 2019; Franco-Lopez et al., 2001; Gou et al., 2019; Thierry et al., 2019; Burges et al., 1998; Chapelle et al., 1999; Janik et al., 2006; Torkaman et al., 2015) provides novel, efficient and economical tools for reservoir engineering and has been proved to be a powerful tool in production forecast. Some researchers use big data technology to build proxy model by correlating the complex, non-linear relationship among parameters to forecast flow rates and hydrocarbon recoveries (Panja et al., 2017; Zhong et al., 2020; Ng et al., 2021; Li et al., 2021; Shen et al., 2022; Zha et al., 2021; Zhou et al., 2014), and in some literatures, big data technology have been utilized to facilitate reservoir simulation in saving run time and cost, or improving accuracy in history matching (Ke et al., 2017; Cheng et al., 2019; Luciana et al., 2020; Feng et al., 2019), and some researchers use big data technology to guide stimulation design by correlating the fracturing parameters into post stimulation oil production prediction model (Zhu et al., 2015).

But less literature is presented to forecast overall DI for a raw gas reservoir with big data analysis technique. The purpose of this paper lies in the point that how we utilized the geology and production history data in developed reservoirs to facilitate the exploitation of new findings. Geology, dynamic and DI data from 30 mature gas fields are collected and processed, and 13 parameters are selected to represent geological features, deliverability and DI of individual reservoir. Through BP Neural Network, proxy models are established to correlate DI with geology and deliverability data. Moreover, the stability of the predicted results are also considered, to avoid randomness in a single experiment, Bagging method (Eugene et al., 2022) is used to make the results more stable for cases with limited samples. With the established models, overall DI for QX reservoir are then given based on current drilling and testing information, and risks caused by heterogeneity are also discussed. The results can serve as a criteria for directing the successful development of this ultra-deep marginal pools.

## 2 Data Acquisition and Processing

Geology, deliverability and DI data of 30 major mature gas reservoirs from different gas-bearing basins in China are reviewed. With per capita porosity among 3.4%–28.6% and per capita dynamic permeability ($K_{\text{dynamic}}$, permeability from well test interpretation) ranging from 0.1–38.5 mD, these reservoirs contain sandstones and carbonate rocks with or without natural fractures. Based on post FDP implementation evaluation of DI, these reservoirs, with 15–691 $10^9$ $m^3$ in OGIP and 0.3–10.7 $10^9$ $m^3$/a in actual plateau gas production, are all believed to be successfully developed reservoirs. In data preparation, logging and dynamic data of $1500^{+}$ wells in these reservoirs are reviewed to better understand the productivity and its dominating factors of individual reservoir. And 13 parameters are selected and listed in Table 1 to represent geological features, deliverability and DI of individual reservoir. To make sure that these parameters fully represent reservoir characteristics, a lot of reservoir engineering study are conducted, especially in the selection of productivity related parameters, such as permeability and well AOF. We have two sets of reservoir permeability, which are matrix permeability ($K_{matrix}$) obtained in well logging interpretations or core testing and dynamic permeability ($K_{dynamic}$) calculated in well test interpretation. The correlations of permeability (both $K_{matrix}$ and $K_{dynamic}$) *vs* porosity, and $K_{dynamic}$ *vs* $K_{matrix}$ shown in Fig. 2 indicate that the porosity for most reservoirs are quite low ($<10\%$), but the permeability varies considerably, and inconsistency exists between $K_{dynamic}$ and $K_{matrix}$ due to the development of natural fractures. Figure 3 indicates that one of the DI parameters—FPR is more dependent on $K_{dynamic}$ than $K_{matrix}$, and it can be seen in Fig. 4 that $K_{dynamic}$ also dominate well deliverability (AOFP).

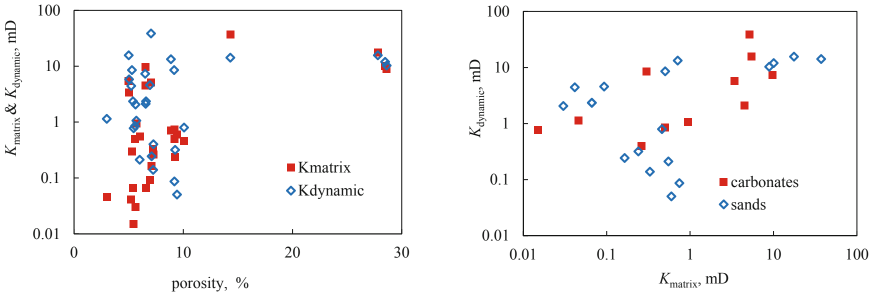
**Table 1.** Geology, deliverability and DI parameters for individual reservoir

| Parameter type | No. of parameters | parameters | scope of values in 30 reservoir samples | Values in QX reservoir |
|---|---|---|---|---|
| Geology | 6 | reservoir depth, m | 910–6800 | 7500 |
| | | reservoir pressure, MPa | 9.8–115.5 | 96 |
| | | pressure coefficient, MPa/100 m | 0.85–2.12 | 1.28 |
| | | reserves abundance, $10^9 m^3/km^2$ | 0.1–5.9 | 0.32 |
| | | average porosity, % | 3.4–28.6 | 3.7 |
| | | average $K_{matrix}$, mD | 0.01–37.3 | 0.51 |
| Deliverability | 2 | average $K_{dynamic}$, mD | 0.1–38.5 | 2.0 |
| | | well average AOFP, $10^3 m^3/d$ | 68–9695 | 1420 |

(*continued*)

**Table 1.** (*continued*)

| Parameter type | No. of parameters | parameters | scope of values in 30 reservoir samples | Values in QX reservoir |
|---|---|---|---|---|
| DI | 5 | Field Annual Production Rate (FAPR), % | 0.18–4.11 | 2.5 |
| | | Field Plateau Period (FPP), a | 5–20 | 9–11 |
| | | Well Spacing Density (WSD), km$^2$/well | 0.4–10.5 | 5–6 |
| | | Ultimate Recovery Factor (URF), % | 37.4–75.0 | 62 |
| | | Well Average Daily Production, $10^3$m$^3$/d | 4–1907 | 280–300 |



(a) $K_{matrix}$ and $K_{dymamic}$ *vs.* porostiy

(b) $K_{dymamic}$ *vs.* $K_{matrix}$

**Fig. 2.** $K_{matrix}$, $K_{dymamic}$ and porostiy in 30 sample fields

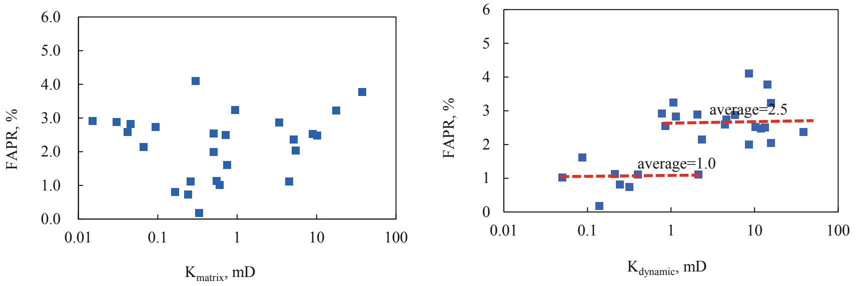


**Fig. 3.** FAPR *vs.* K$_{\text{matrix}}$ and K$_{\text{dynamic}}$ in 30 sample fields

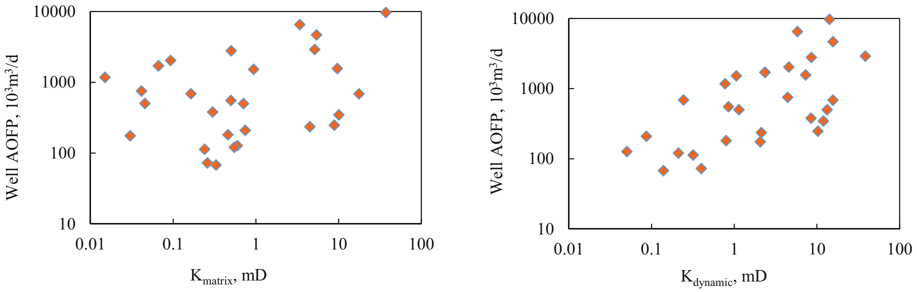


**Fig. 4.** Well AOFP *vs.* K$_{\text{matrix}}$ and K$_{\text{dynamic}}$ in 30 sample fields

## 3 BP Neural Network Algorithms

The BP (Back Propagation) neural network is a data mining technique in developing correlation models between input variables and output variables in big data analysis. In the following we would briefly describe the main algorithms.

The BP neural network, a concept introduced by scientists in 1986, is a multilayer feed-forward neural network trained according to an error back propagation algorithm and is one of the most widely used neural network models (Burks et al., 2000, Meinel et al., 2010). Currently, the vast majority of neural network models used in the practical application of artificial neural networks are in the form of BP networks and variations of it.

The BP algorithm consists of two processes: the forward propagation of the signal and the backward propagation of the error (Hecht-Nielsen, 1989). In forward propagation, the input samples are passed in from the input layer, processed in turn by the hidden layer and then passed to the output layer. If the actual output of the output layer does not match the desired output, it moves to the back propagation of error stage. The BP network consists of an input layer, an output layer and a hidden layer, and the structure of the BP neural network is as follows (Fig. 5).

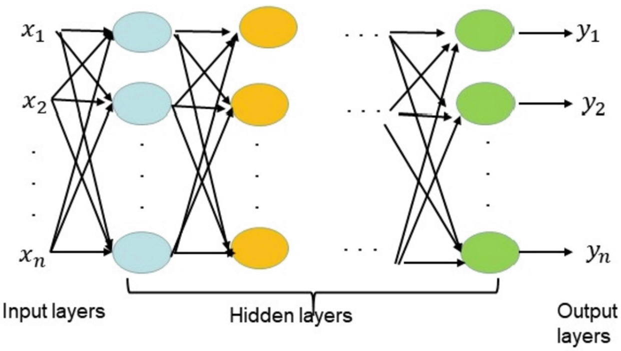


**Fig. 5.** Structure of BP neural network

The specific steps are: let the input vector is $X = (x_1, x_2, \ldots, x_\text{n})$, the input vector of the hidden layer is $hi = (hi_1, hi_2, \ldots, hi_\text{p})$, the output vector of the hidden layer is $ho = (ho_1, ho_2, \ldots, ho_\text{p})$, the input vector of the output layer is $yi = (yi_1, yi_2, \ldots, yi_\text{q})$, the output vector of the output layer is $yo = (yo_1, yo_2, \ldots, yo_\text{q})$, the desired output vector is $d_o = (d_1, d_2, \ldots, d_\text{q})$.

The input and output of each neuron in the hidden layer are calculated by randomly selecting the kth input sample.

$$hi_h(k) = \sum\nolimits_{i=1}^{n} w_{ih}x_i(k) - b_h \quad h = 1,\ 2,\ \ldots,\ p \tag{1}$$

$$ho_h(k) = f(hi_h(k)) \quad h = 1,\ 2,\ \ldots,\ p \tag{2}$$

$$yi_o(k) = \sum\nolimits_{h=1}^{p} w_{ho}ho_h(k) - b_o \quad o = 1,\ 2,\ \ldots,\ q \tag{3}$$

$$yo_o(k) = f(yi_o(k)) \quad o = 1,\ 2,\ \ldots,\ q \tag{4}$$

where $w_\text{ih}$ is the connection weight of the input layer to the middle layer, $w_\text{ho}$ is the connection weight of the hidden layer to the output layer, $b_\text{h}$ is the threshold of each neuron in the hidden layer, and $b_\text{o}$ is the threshold of each neuron in the output layer, $f()$ is the activation function.

Initialize the error function with a random number within $(-1, 1)$ and set the precision $\varepsilon$. With a maximum number of iterations M, the error function is

$$e = \frac{1}{2}\sum\nolimits_{o=1}^{q} (d_o(k) - yo_o(k))^2 \tag{5}$$

Calculate the partial derivatives of the error function with respect to each neuron in the output layer and calculate the parameters of each layer with following equation:

$$\frac{\partial e}{\partial w_{ho}} = \frac{\partial e}{\partial yi_o}\frac{\partial yi_o}{\partial w_{ho}} = \delta_o(k)ho_h(k) \tag{6}$$

$$\begin{aligned}\frac{\partial e}{\partial yi_o} &= \frac{\partial\left(\frac{1}{2}\sum_{o=1}^{q}(d_o(k)-yo_o(k))\right)^2}{\partial i_o}\\ &= -(d_a(k)-yo_o(k))yo_o'(k)\cdot\overline{n}e(d_o(k)-yo_o(k))f(yi_o(k))=\delta_o(k)\end{aligned} \tag{7}$$

$$\frac{\partial yi_o(k)}{\partial w_{ho}} = \frac{\partial\left(\sum_h^p w_h ho_h(k)-b_o\right)}{\partial w_{ho}} = ho_h(k) \tag{8}$$

Calculate the partial derivatives of the error function for each neuron in the hidden layer, the connection weights that follow, and the input values for that layer,

$$\frac{\partial e}{\partial hi_h(k)} = -\left(\sum_{h=0}^{q}\delta_0(k)w_{ho}\right)f'(hi_h(k)) = \delta_h(k) \tag{9}$$

$$\frac{\partial hu_h(k)}{\partial w_{ih}} = x_i(k) \tag{10}$$

$$\frac{\partial e}{\partial w_{ih}} = \delta_h(k)x_i(k) \tag{11}$$

Use (6) (7) (8) to correct the output layer connection weights,

$$\Delta w_{ho}(k) = -\mu\frac{\partial e}{\partial w_{ho}} = \mu\delta_o(k)ho_h(k) \tag{12}$$

$$w_{ho}^{N+1} = w_{ho}^{N} + \eta\delta_o(k)ho_h(k) \tag{13}$$

Use (9) (10) (11) to correct the hidden layer connection weights,

$$\Delta w_{ih}(k) = -\mu\frac{\partial e}{\partial w_{ih}} = -\mu\frac{\partial e}{\partial hi_h(k)}\frac{\partial hi_h(k)}{\partial w_{ih}} = \delta_h(k)x_i(k) \tag{14}$$

$$w_{ih}^{N+1} = w_{ih}^{N} + \eta\delta_h(k)x_i(k) \tag{15}$$

Finally, calculate the global error,

$$E = \frac{1}{2}\sum_{k=1}^{m}\sum_{o=1}^{q}(d_o(k)-y_o(k))^2 \tag{16}$$

## 4 Bagging

Bagging is a parallel method of ensemble learning, where data is sampled and the results are voted on. For a given data set containing multiple samples, we randomly select one sample into the sampling set and put that sample back into the initial data set, making it still possible for it to be selected for the next sampling. Combining Bagging with BP neural network. The model is trained several times to get the average of the predicted values. It can improve the accuracy and stability of prediction while avoiding the over-fitting phenomenon.

# 5 DI Prediction Proxy Model Development Through Big Data Analysis

## 5.1 Correlation Coefficient Calculation

As we want to build DI prediction model through big data analysis, geology and deliverability parameters listed in Table 1 are categorized as characteristic data and DI parameters in the table are defined as target output variables. In proxy model development, initially, the coefficient of correlation *r* (Bookbinder et al., 1987) between characteristic data and target output data are calculated and those characteristic data with high absolute *r* values are selected as input parameters. Table 2 presents the calculated *r* values between characteristics data and DI, and those characteristic data with underlined values are selected as inputs.

**Table 2.** Calculated *r* values between Characteristic Data and DI

| DI parameters | Depth | Pressure | Pressure Coefficient | Reserves Abundance | $K_{\text{matrix}}$ | Porosity | $K_{\text{dynamic}}$ | AOFP |
|---|---|---|---|---|---|---|---|---|
| URF | **0.358** | **0.333** | 0.204 | 0.298 | 0.273 | −0.176 | **0.401** | **0.522** |
| FPP | −0.006 | **0.146** | **0.247** | 0.065 | **0.168** | 0.139 | −0.127 | 0.136 |
| FAPR | −0.023 | 0.12F9 | 0.334 | **0.385** | 0.190 | 0.204 | **0.392** | **0.416** |
| WSD | 0.223 | 0.225 | 0.101 | −**0.318** | -0.091 | −**0.450** | 0.123 | 0.167 |
| WADP | 0.326 | 0.494 | 0.624 | **0.735** | **0.745** | −0.007 | 0.388 | **0.948** |

*Note:those characteristic data with underlined values are selected as inputs*

## 5.2 Proxy Model Development

In proxy model development, BP Neural Network is used to establish the relationship between input variables and output variables. We design different combinations of correlated variables as input models. For example, we design four input models for predicting UFR and three input models for predicting FPR, as shown in Tables 3 and 4 respectively.

**Table 3.** Combinations of correlated variables as inputs for predicting URF

| Input models | Depth | Pressure | $K_{\text{dynamic}}$ | AOFP |
|---|---|---|---|---|
| Model 1 | 1 | 1 | 1 | 1 |
| Model 2 | 1 | 0 | 1 | 1 |
| Model 3 | 0 | 1 | 1 | 1 |
| Model 4 | 0 | 0 | 1 | 1 |

*Note:1 means the variable is used, 0 means it is not used*

**Table 4.** Combinations of correlated variables as inputs for predicting FAPR from gas fields

| Input models | Reserves Abundance | $K_{dynamic}$ | AOFP |
|---|---|---|---|
| Model 1 | 1 | 1 | 1 |
| Model 2 | 0 | 1 | 1 |
| Model 3 | 1 | 0 | 1 |

*Note:1 means the variable is used, 0 means it is not used*

As limited sample data may introduce randomness and occasionality in model development, thus weaken model credibility, to avoid these disadvantages, samples data are disordered in each training with 80% and 20% being selected randomly for model training and model validating respectively. For fixed inputs and outputs, the risks of occasionality caused by limited sample data also exist if only single numerical test is conducted, to tackle this problem, 2000 numerical tests are performed and those models with high coefficient of determination ($R^2$) of test set are selected as best fit models. All best fit models are used to predict the DI value, and then the average is calculated to obtain the final prediction result. $R^2$ is generally used in regression models to evaluate the degree of conformity between predicted values and actual values, and $R^2$ is defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

where: $\overline{y}$ denotes the average of the true target values. The higher the score of the $R^2$, the closer the predicted value of the sample is to the true value.

In summary, we calculate the correlation coefficients between the predictor variables and other variables, find the most relevant variables to the predictor variables. Then use the bagging-based BP Neural Network to establish the relationship between the predictor variables and the relevant variables. Finally, the training effect is evaluated by $R^2$. And the final prediction value is obtained by averaging from the better prediction results. Whole process of the algorithm is described in the Appendix, and Table 5 shows the network parameter settings for WADP prediction experiments.

**Table 5.** Optimal values for each parameter in the predicted WADP

| Parameters | Setting |
|---|---|
| Units | 10, 64, 128, 256 …, 64, 1 |
| Epochs | Period: 500 |
| activation | Relu |
| optimizer | Adam |

Figure 6 shows, as an example, the prediction results of the two experiments with higher $R^2$ in 2000 prediction experiments of WADP. As depicted in Fig. 6, sound fitting can be observed between prediction and actual values. The predicted value in Fig. 6(a) is basically consistent with the measured value, while the predicted value in Fig. 6(b) is slightly deviated from the measured value, but the error is still small in the case of a small amount of data. Experimental results show that the bagging-based BP neural network has high precision in DI prediction. In addition, this analysis method is easily scalable with the addition of the latest machine learning methods.

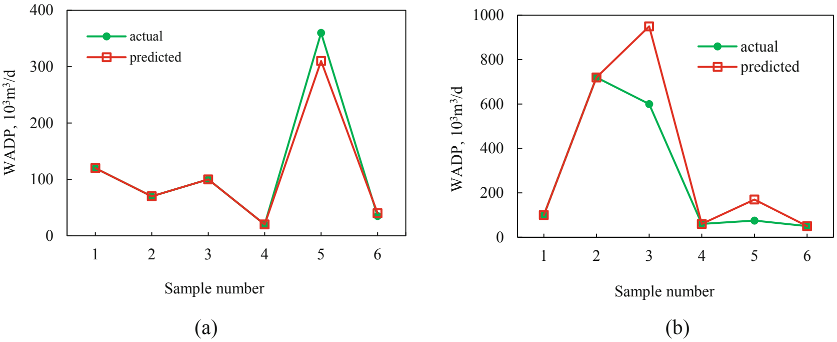


**Fig. 6.** WADP prediction models validation. (a) and (b) represent different models with different numerical testing samples

### 5.3 DI Prediction for QX Reservoir

Current drilling, geology and well testing data in QX reservoir are reviewed, and characteristic parameters including geology and deliverability data are evaluated based on our understanding of the reservoir. The quantifying of these parameters will be discuss below and their values are presented in Table 1. Reservoir mid-depth based on drilling wells is 7500m with initial reservoir pressure of 96 MPa and pressure gradient 1.28 MPa/100 m; reservoir porosity from both core analysis and logging interpretations are among 2.0%–6.0%, with 3.7% in average; $K_{\text{matrix}}$ from core analysis range from 0.01 mD to 53 mD, and 0.51 mD in mean; $K_{\text{dynamic}}$ obtained through 9 wells' test interpretations are ranging in the scope of 0.1–10 mD, with 2.0 mD in average, reflecting the improvement of mobility with the development of natural fractures; AOFP from both horizontal wells and vertical wells are among 0.15–3.85 $10^6$ $m^3$/d, with average 1.42 $10^6$ $m^3$/d; based on logging and pressure data, average reserve abundance is evaluated as 0.32 $10^9$ $m^3/km^2$.

DI for QX reservoir are predicted through our proxy models with input parameters given in Table 1, and the output results shown in Table 1 are as follow: FAPR 2.5%, WSD 5–6 $km^2$/well, WADP during FPP 280–300 $10^3$ $m^3$/d and URF 62%. Heterogeneity caused by lithology change or uneven development of natural fractures can be evidenced from both core samples and deliverability data, as in the low part of structure, well dynamic permeability are in the magnitude of 0.1mD. The influence of heterogeneity on FAPR and URF are also predicted in the proxy models, and results presented in Fig. 7 show that in the "tight" part of the reservoir, the feasible FAPR decreases from 2.5% to

1.5%, and URF declines from 62% to 50%, so economic risk exist in the development of QX reservoir. The effects of horizontal drilling on FAPR and URF are also evaluated and depicted in Fig. 7, and it can be seen that compared with vertical drilling (with average AOFP 1.0MMm$^3$/d), horizontal drilling (with average AOFP 1.4MMm$^3$/d) show limited enhancement in both FAPR and URF.

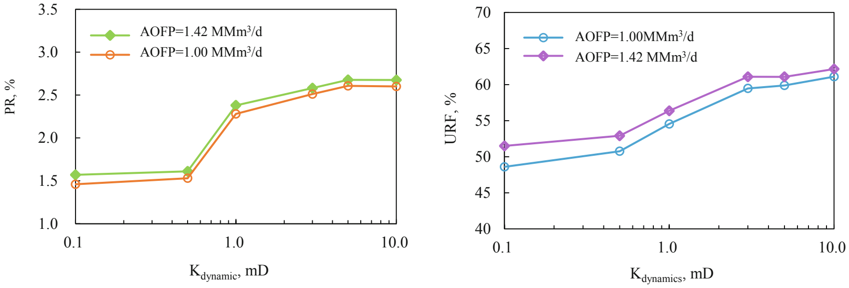


**Fig. 7.** Predicted *PR* and *URF vs.* $K_{\text{dynamic}}$ with different *AOFP*

It should be note the proxy model are based on data from 30 mature, successfully developed major reservoirs, and the DI of these reservoirs contain certain development policies followed currently by the operators. So the predicted DI for QX reservoir can serve as a criteria for directing the successful development of this ultra-deep marginal pools.

## 6 Conclusions

1. DI prediction models for raw gas reservoirs are established through big data analysis approach. Geology and dynamic data from 30 mature gas reservoirs are reviewed, and 12 parameters are selected to represent geology, deliverability and DI data for individual reservoir, then proxy model are built through bagging-based BP neural network to correlated DI with geology and deliverability data.
2. Experimental results show that the bagging-based BP neural network has high precision in DI prediction in the case of limited sample data.
3. Based on geology and dynamic data, the DI for QX reservoir are predicted in the proxy model with results as following: FAPR 2.5%, FPP 9–11 years, URF 62%, WSD 5–6 km$^2$/well, WADP during FPP 280–300 10$^3$ m$^3$/d. Sensitivity analysis showed that for relative "tight" area, 1.5% of FAPR with UFR of 50% are expected.
4. Through big data analysis, the development polices formed in mature gas fields can provide valuable knowledge in the development of ultra-deep raw gas fields, thus mitigating risks due to uncertainties in reservoir characterization.

**Funding.** This study was supported by the Scientific and Technology Research Program Funded by CNPC, China (Project No. 2021DJ1505 and 2022KT0905).

## Appendix

### Algorithm for Prediction Model Devlopment

```
Input: Select variable combination models as input
Initialize training data and set test number
repeat
    $k \leftarrow k+1$
    for $j=1$ to $N$ do in parallel
        $\hat{y}_j \leftarrow M_j(X,y)$
        Calculate $R^2$ value $c_j$, get set $S_j=\{\hat{y}_j,c_j\}$
    end for
    Select $y_s \in S_j$ where $j$ corresponds to $c_j \geq Average(\sum_{j=1}^{N} c_j)$
until $k=K$
    $\hat{y}=Average(\sum y_s)$
return $\hat{y}$
```

## References

Bookbinder, M.J., Panosian, K.J.: Using the coefficient of correlation in method-comparison studies. Clin. Chem. **7**, 1170–1176 (1987)

Burges, C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Disc. **2**(2), 121–167 (1998)

Burks, T.F., Shearer, S.A., Gates, R.S., et al.: Backpropagation neural network design and evaluation for classifying weed species using color image texture. Trans. Asae **43**(4), 1029–1037 (2000)

Chapelle, O., Haffner, P., Vapnik, V.N.: Support vector machines for histogram-based image classification. IEEE Trans. Neural Netw. **10**(5), 1055–1064 (1999)

Cheng, Z., Sankaran, S., Lemoine, V., et al.: Application of machine learning for production forecasting for unconventional resources. Paper URTeC-2019-47 Presented at Unconventional Resources Technology Conference, Colorado (2019)

Eugene, L., Chieh-Hsin, L., Hsien-Yuan, L.: A bagging ensemble machine learning framework to predict overall cognitive function of schizophrenia patients with cognitive domains and tests. Asian J. Psychiatr. **69**, 103008 (2022)

Feng, C., Li, J., Feng, Z., et al.: Predict oil production from geological and petrophysical data before hydraulic fracturing using an improved particle swarm optimization based least squares support vector machine. Paper SPE 197250 Presented at Abu Dhabi International Petroleum Exhibition & Conference (2019)

Franco-Lopez, H., Ek, A.R., Bauer, M.E.: Estimation and mapping of forest stand density, volume and cover type using the k-nearest neighbors method. Remote Sens. Environ. **77**, 251–274 (2001)

Gou, J.P., Ma, H.X., Ou, W.H., et al.: A generalized mean distance-based k-nearest neighbor classifier. Expert Syst. Appl. **115**(1), 356–372 (2019)

Hecht-Nielsen, R.: Theory of the backpropagation neural network. Neural Netw. (1989)

Janik, P., Lobos, T.: Automated classification of power-quality disturbances using SVM and RBF networks. IEEE Trans. Power Deliv. **21**(3), 1663–1669 (2006)

Ke, G.L., Meng, Q., Thomas, F., et al.: Light GBM: a highly efficient gradient boosting decision tree. Neural Inf. Process. Syst. (2017)

Li, D.L., Shen, L.H., Zha, W.S., et al.: Physics-constrained deep learning for solving seepage equation. J. Petrol. Sci. Eng. **206**, 1–11 (2021)

Luciana, M.D.S., Guilherme, D.A., Denis, J.S.: Support vector regression for petroleum reservoir production forecast considering geostatistical realizations. SPE Reservoir Eval. Eng. **23**(04), 1343–1357 (2020)

Meinel, L.A., Stolpen, A.H., Berbaum, K.S., et al.: Breast MRI lesion classification: improved performance of human readers with a backpropagation neural network computer-aided diagnosis (CAD) system. J. Magn. Reson. Imaging **25**(1), 89–95 (2010)

Ng, C.S.W., Ghahfarokhi, A.J., Amar, M.N.: Well production forecast in Volve field: application of rigorous machine learning techniques and metaheuristic algorithm. J. Petrol. Sci. Eng. **208**, 109468 (2021)

Panja, P., Velasco, R., Pathak, M., et al.: Application of artificial intelligence to forecast hydrocarbon production from shales. Petroleum 75–89 (2017)

Quinlan, J.R.: Induction of decision trees. Mach. Learn. **1**(1), 81–106 (1986)

Rao, H.D., Shi, X.Z., Rodrigue, A.K., et al.: Feature selection based on artificial bee colony and gradient boosting decision tree. Appl. Soft Comput. **74**, 634–642 (2019)

Safavian, S.R., Landgrebe, D.: A survey of decision tree classifier methodology. IEEE Trans. Syst. Man Cybern. **21**(3), 660–674 (1991)

Shen, L.H., Li, D.L., Zha, W.S., et al.: Surrogate modeling for porous flow using deep neural networks. J. Petrol. Sci. Eng. **213**, 110460 (2022)

Thierry, D., Orakanya, K., Songsak, S.: A new evidential K-nearest neighbor rule based on contextual discounting with partially supervised learning. Int. J. Approximate Reasoning **113** (2019)

Torkaman, M., Safari, et al.: A novel PSO-LSSVM model for predicting liquid rate of two phase flow through wellhead chokes. J. Natural Gas Sci. Eng. **24**, 228–237 (2015)

Zha, W.S., Zhang, W., Li, D.L., et al.: Convolution-based model-solving method for three-dimensional, unsteady, partial differential equations. Neural Comput. 1–23 (2021)

Zhong, Z., Sun, A.Y., Wang, Y., et al.: Predicting field production rates for waterflooding using a machine learning-based proxy model. J. Petrol. Sci. Eng. **194**, 107574 (2020)

Zhou, Q.M., Robert, D., Andrew, K., et al.: Evaluating gas production performances in Marcellus using data mining technologies. J. Nat. Gas Sci. Eng. **20**, 109–120 (2014)

Zhu, L., Li, M.S., Wu, Q.H., et al.: Short-term natural gas demand prediction based on support vector regression with false neighbours filtered. Energy **80**(2), 428–436 (2015)

# Calibration Technology and Application of Mud Logging Sensors Based on Artificial Intelligence

Chang-liang Wu[1,2,3](✉), Zhi-xiong Zhou[1,2,3], Tie-heng Ding[1,2,3], Jian-guo Xiong[1,2,3], Yong-liang Gao[1,2,3], Yang Li[1,2,3], and Xue-li Luo[1,2,3]

[1] KEMBL Petroleum Technology Co., Ltd., Beijing, China
wuchangliangdr@cnpc.com.cn
[2] CNPC Engineering Technology R&D Company Ltd., Beijing, China
[3] National Engineering Research Center of Oil and Gas Drilling and Completion Technology, Beijing, China

**Abstract.** Mud logging serves as the "eyes" of exploration and development, acting as a counselor for drilling safety, the center of information transmission, and holding the first-hand data on oil and gas exploration and development. With the rapid development of informatization, digitization, intelligence, and remote support systems, the demand for high-quality mud logging data has continuously risen, where sensor calibration and calibration technology serve as the foundation for ensuring accuracy and reliability. This paper proposes an artificial intelligence-based comprehensive mud logging instrument sensor calibration and calibration technology, targeting the issues of prolonged service life, low precision, and low inspection rate of traditional mud logging instruments. The technology primarily involves collecting and pre-processing sensor output data such as filtering, sampling to eliminate noise, and improve the dataset's quality. Mathematical models of sensors were constructed using machine learning or deep learning algorithms to analyze the relationship between sensor outputs and actual values, which could also compute sensor errors and uncertainties. Algorithm optimization methods such as wavelet transform and adaptive filtering were used to process and analyze sensor data for different types of sensors and environmental conditions. The adaptive control algorithm was then utilized based on the predicted model results and actual measurement results to calibrate the sensor, ultimately helping to avoid

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20-22 September 2023.

This paper was selected for presentation by the IFEDC Committee following review of information contained in an abstract submitted by the author(s). Contents of the paper, as presented, have not been reviewed by the IFEDC Technical Team and are subject to correction by the author(s). The material does not necessarily reflect any position of the IFEDC Technical Committee its members. Papers presented at the Conference are subject to publication review by Professional Team of IFEDC Technical Committee. Electronic reproduction, distribution, or storage of any part of this paper for commercial purposes without the written consent of IFEDC Organizing Committee is prohibited. Permission to reproduce in print is restricted to an abstract of not more than 300 words; illustrations may not be copied. The abstract must contain conspicuous acknowledgment of IFEDC. Contact email: paper@ifedc.org.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 113–123, 2024.
https://doi.org/10.1007/978-981-97-0272-5_9

errors and uncertainty in the traditional manual calibration process. Experimental results show that this technology has higher accuracy and reliability than traditional calibration techniques while maintaining simple operation, fast speed, and cost-effectiveness. This technology improves the level of detection and evaluation technology of comprehensive mud logging instruments, Standardizes mud logging equipment management, and plays an essential role in timely discovering, evaluating oil and gas layers, and optimizing drilling construction safety.

**Keywords:** Artificial Intelligence · Mud Logging · Sensors · Calibration Technology

## 1 Introduction

Informationization, digitization, intelligentization, and remote support systems require high-quality mud logging data. The comprehensive mud logging instrument is the main technical equipment on-site for mud logging, responsible for data collection, processing, analysis, and transmission. It monitors engineering parameters and drilling fluid parameters in real-time during the drilling process, analyzes various gas contents in the drilling fluid. The accuracy and reliability of mud logging data directly affect the quality and safety of drilling projects. It is the basis for timely discovering and evaluating oil and gas layers and optimizing drilling construction safety. Currently, there are several factors that affect the quality of mud logging data.

(1) The harsh installation and usage conditions of the mud logging sensors may cause reduced accuracy, malfunctions, and damages (as shown in Fig. 1).
(2) Mechanical vibration and impact caused by frequent lifting and long-distance transportation can damage equipment in the instrument room. Chipsets and electronic components will have degraded performance as their service time increases, which can lead to abnormal data channels or reduced conversion accuracy.
(3) The performance of the gas analysis system will decrease with production and operation time.
(4) Similar to drilling operations, mud logging operations are located in remote locations with difficult-to-control environmental conditions. Existing indoor testing and assessment devices have low integration and large size and weight, which cannot meet on-site testing needs, resulting in delayed and incomplete testing and assessment.
(5) Comprehensive mud logging instruments of different brands and periods have significant differences in quality, configuration, performance, etc. Many instruments have been in service for more than 10 years.
(6) Most mud logging companies mainly focus on individual testing, calibration, and verification, lacking a unified and systematic comprehensive mud logging instrument testing and evaluation device and technical specifications.

Therefore, major petroleum companies at home and abroad attach great importance to mud logging work, regarding improving mud logging equipment performance and ensuring mud logging data quality as the basis for improving mud logging quality and engineering technical data quality. In order to eliminate the impact of the above unfavorable factors, research on comprehensive mud logging instrument testing and evaluation technology and equipment has been conducted, forming a set of complete technical specifications for comprehensive mud logging instrument testing and evaluation, and developing a comprehensive mud logging instrument testing and evaluation device that can adapt to fieldwork.

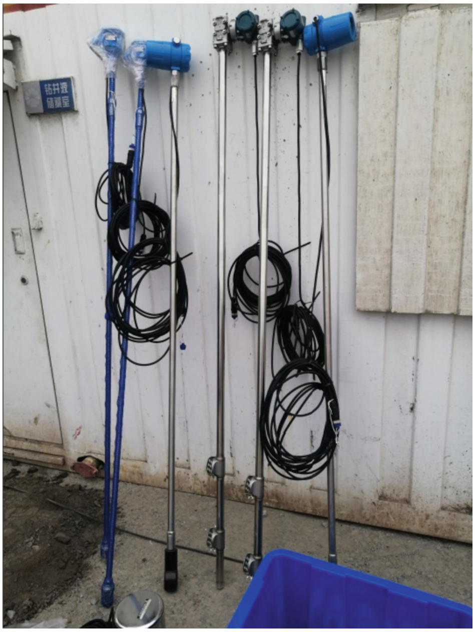


**Fig. 1.** Partial sensors of comprehensive mud logging instrument

Significant progress has been made in recent years with the application of artificial intelligence technology. Many researchers apply AI to sensor-based health and sports biomechanics [1–7], while others utilize it for intelligent industrial manufacturing [8–10]. In the area of using artificial intelligence for sensor calibration, many scholars have also conducted extensive research and achieved significant progress [11–16].

## 2 Performance Testing, Data Acquisition and Preprocessing of Sensor

### 2.1 Performance Testing of the Sensors

Sensor performance testing and data acquisition and preprocessing form the foundation of artificial intelligence-based calibration technology for comprehensive logging instrument sensors. A complete set of sensor performance testing equipment was developed using a Siemens 16-bit high-precision PLC, high-precision pressure source, high and low-temperature constant temperature water tank, standard current generator, and supporting coils and high-precision resistors, and an accompanying system was programmed in C#.

Sensor performance testing: Standard testing equipment is used to test the performance of the logging sensors and record measurement results and error data. Sensor performance testing usually includes the following indicators: sensitivity, resolution, accuracy, and response time. The performance of the sensor is determined by testing it through methods such as adding a known quantity to the sensor or directly placing it in a changing environment and collecting feedback signals. After completing the sensor performance testing, the sensor is evaluated against specific application requirements. In short, sensor performance testing is an important step in ensuring data accuracy and reliability.

## 2.2 Data Collection and Preprocessing

The data collection system is used to collect the data obtained from the above tests and preprocess it for feature extraction by machine learning algorithms. Different signal acquisition methods, either analog or digital, are employed depending on the type of sensor. Then, the collected data must be preprocessed to remove noise, artifacts, and other unwanted signals.

Filtering is a common data preprocessing technique that can separate useful signals from noisy signals by applying filters. Depending on the filtering method, it can be classified into various types such as low-pass filtering, high-pass filtering, and band pass filtering. Sampling refers to the process of discretizing raw data by converting continuous analog signals into discrete digital signals, making them easier to store and process.

Filtering is a process that removes or retains certain components of a signal via filters. Its mathematical principles are based on signal processing theory. Common filtering methods include moving average filtering, median filtering, IIR low-pass filtering, FIR low-pass filtering, and frequency domain filtering. For logging sensor signal filtering, FIR low-pass filtering is used.

The mathematical formula for FIR low-pass filter can be expressed as:

$$y(n) = \sum\nolimits_{k=0}^{M} h(k)x(n-k) \tag{1}$$

where $x(n)$ represents the original signal, $y(n)$ represents the filtered signal, and $h(k)$ represents the coefficients of the filter. The order of the filter is denoted by $M$.

A linear phase FIR filter is adopted, and the specific formula for calculating its coefficients is as follows:

$$h(k) = \begin{cases} 2f_c, k = \frac{M}{2} \\ \frac{\sin(2\pi f_c\left(k-\frac{M}{2}\right))}{\pi\left(k-\frac{M}{2}\right)}, k \neq \frac{M}{2} \end{cases} \tag{2}$$

In the above formula, $h(k)$ represents the coefficients of the filter, $M$ is the order of the filter, and $f_c$ is the cutoff frequency of the filter. After calculating the filter coefficients using the aforementioned formula, they can be applied to convolution operations to filter the original signal and obtain the filtered results.

After applying the standard excitation signal generated by the high-precision standard source to the logging sensor, an analog current signal will be generated by the sensor, which is generally located between 4–20 mA. With the developed signal acquisition instrument, the current signal can be read and converted into corresponding physical quantities to complete the calibration of the sensor. Figure 2 shows the sensor signal acquisition instrument, and Fig. 3 shows the working of the electric torque sensor calibration device.

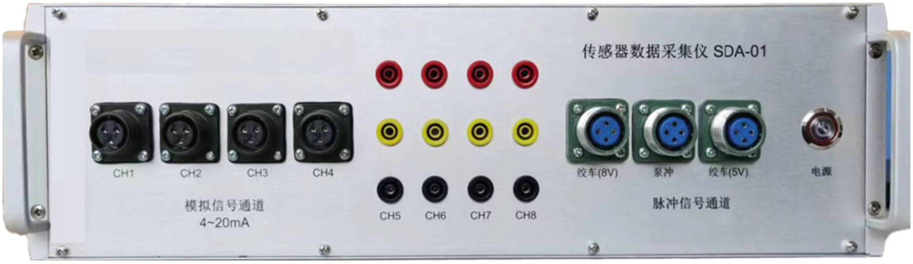


**Fig. 2.** SDA-01 Sensor Data Acquisition Instrument

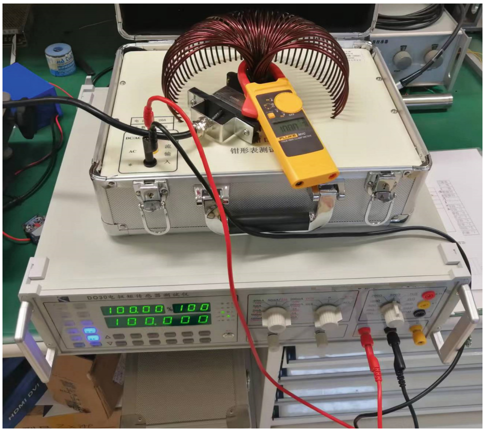
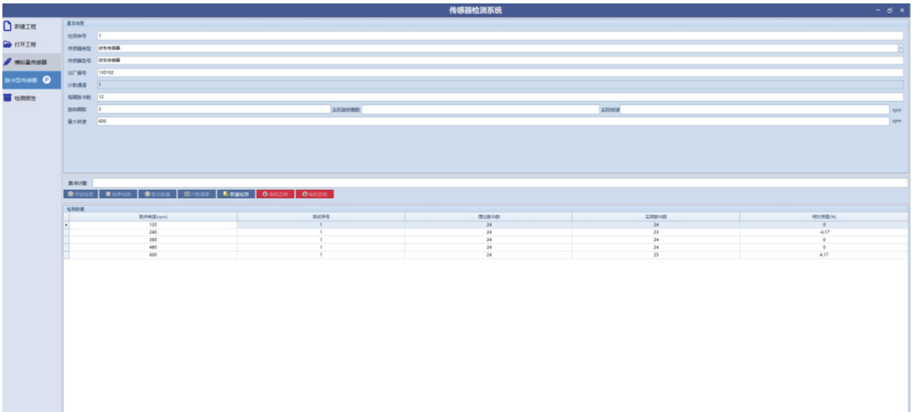


**Fig. 3.** Calibration of electric torque sensors



基本信息

检测序号 1

传感器类型 绞车传感器

传感器型号 绞车传感器

出厂编号 130102

计数通道 1

每圈脉冲数 12

旋转圈数 2 实际旋转圈数 实际转速 rpm

最大转速 600 rpm

脉冲计数

检测数据

| 测试转速(rpm) | 测试序号 | 理论脉冲数 | 实测脉冲数 | 相对误差(%) |
|---|---|---|---|---|
| 120 | 1 | 24 | 24 | 0 |
| 240 | 1 | 24 | 23 | -4.17 |
| 360 | 1 | 24 | 24 | 0 |
| 480 | 1 | 24 | 24 | 0 |
| 600 | 1 | 24 | 25 | 4.17 |

**Fig. 4.** Interface of the calibration system for comprehensive logging instrument sensors

## 3 Model Establishment Based on Bayesian Optimization

After obtaining the raw data of mud logging sensors, the optimal model structure and hyperparameter combination are searched through Bayesian optimization to obtain a better theoretical curve.

Bayesian optimization is a black-box function optimization method commonly used in scenarios where a target function needs to be maximized or minimized. When constructing a Bayesian optimization model, we need to define a Gaussian process to describe the overall trend and uncertainty information of the target function. We also need to define a surrogate function to approximate the target function and optimize the surrogate function to find the optimal solution of the target function.

Specifically, the following steps are taken to construct the Bayesian optimization model:

(1) Define the prior distribution of the Gaussian process. In this step, we need to define a mean function and a covariance function for the Gaussian process. The mean function is used to describe the average value of the target function at different input values, while the covariance function is used to describe the correlation between different input values. The typically chosen Gaussian process prior distribution is the zero-mean Gaussian process.

(2) Update the posterior distribution of the Gaussian process based on the existing data. In this step, we need to update the mean function and covariance function of the Gaussian process based on the existing sample data to obtain a more accurate function approximation.

(3) Calculate the next sampling point based on the surrogate function. In this step, we need to use the current Gaussian process to fit the target function, construct a surrogate function, and select the next sampling point by optimizing the surrogate function. Common optimization methods include greedy algorithm and coordinate axis optimization.

(4) Update the posterior distribution of the Gaussian process based on the new sampling point. After obtaining the new sampling point, it can be added to the existing samples,

and these data can be used to update the mean function and covariance function of the Gaussian process.

(5) Repeat steps 3 and 4 until the preset stopping conditions are met.

The entire process of Bayesian optimization can be mathematically expressed as follows:

$$x_{i+1} = arg\max_{x} EI(x|D_t) = arg\max_{x}(\mu_t(x) - \xi\sigma_t(x) \tag{3}$$

In the equation, $x_{\text{t+1}}$ represents the next sampling point chosen in the iteration, $EI(x|D_t)$ is the expected improvement metric, representing the expected increase in target function value over the current best known value, given x as input under Gaussian process fitting. $\mu_t(x)$ and $\sigma_t(x)$ represent the mean and standard deviation of the current Gaussian process at x, respectively. $\xi$ is a hyperparameter that controls the balance between exploration and exploitation and is commonly set to 2 or 3".

# 4 Model Training and Experimental Verification

## 4.1 Model Training

Neural networks are models composed of neurons that utilize components such as weights, biases, and activation functions to facilitate information transmission and processing. These models possess remarkable fitting and expressive abilities, making them suitable for solving various machine learning and deep learning tasks. Therefore, in this study, the neural networks were used to train the sensor calibration data. Simultaneously, the Cross Entropy loss function and the Stochastic Gradient Descent (SGD) optimizer was selected as key components of the neural network and combined to train a more accurate and efficient model.

Model Training: Using a large-scale dataset to train the model, constantly updating the model parameters to improve prediction accuracy and robust performance.

To train the model with a large-scale dataset, the following steps are required:

(1) Data collection and preparation: First, it is necessary to obtain enough data to train the model, and the data should be representative and able to cover various possible situations. Then, the data needs to be cleaned, transformed, and normalized, so that the model can better understand and process it.

(2) Model selection and design: Based on the application scenario and data characteristics, select an appropriate model structure and determine the parameters that need to be optimized.

(3) Loss function and optimizer selection: Depending on the task of the model, select an appropriate loss function and optimizer to evaluate the model and adjust the model parameters.

(4) Batch training: Since the dataset is too large to be loaded into memory for training at once, the data needs to be divided into equally-sized batches, and the stochastic gradient descent algorithm (SGD) is used to update the model parameters batch by batch.

(5) Batch normalization and regularization: Performing batch normalization before or after each batch can reduce the bias and variance of input features, thereby improving the model's prediction accuracy and robustness. In addition, methods such as L1 or L2 regularization can constrain the size and number of model parameters, avoiding overfitting and underfitting.

(6) Model evaluation and fine-tuning: Evaluate the model through the training and testing sets to determine the model's prediction accuracy and robust performance. If problems are found in the model, fine-tuning is needed, such as changing the model structure, adjusting the loss function or optimizer, etc.

Through these steps, the model can be trained using a large-scale dataset, constantly updating the model parameters to improve prediction accuracy and robust performance.

When choosing the appropriate network architecture, number of layers, and number of nodes to establish the ANN model and initialize weights, several steps usually need to be performed:

The problem type is determined by firstly clarifying whether a classification problem or a regression problem is faced. This will help determine the network structure and activation function.

Input and output are determined by specifying the number and type of input feature vectors and output predicted values.

An appropriate activation function is chosen based on the problem type, such as sigmoid or ReLU.

The network structure is designed by selecting a network structure that includes determining the range of the number of nodes in each hidden layer, whether to use dropout techniques, and so on.

Weights are initialized by selecting appropriate weight initial values, such as Xavier initialization, etc.

The model is trained by using the training dataset, and parameters are adjusted based on the validation set results.

The model is evaluated by examining its performance using the testing set.

When selecting the network structure and number of nodes, the principle of Occam's Razor should be followed. That is to say, the model structure should be made as simple as possible with reduced node numbers to prevent overfitting. At the same time, when designing the model, common deep learning frameworks such as TensorFlow and PyTorch can be considered. They provide a series of optimized structures, numbers of layers, and nodes, as well as pre-trained weights, which can reduce some manual parameter tuning work.

Taking the casing pressure sensor as an example, with a measuring range of 0 ~ 70MPa, it should be calibrated using a standard pressure pump source calibrated by the Beijing Institute of Metrology and Measurement. The training samples are shown in the table below.

**Table 1.** Training sample for calibration of casing pressure sensor.

| Input | Output | Result |
|---|---|---|
| 0MPa | 4.00 mA | Pass the calibration |
| 0MPa | 4.01 mA | Pass the calibration |
| 0MPa | 4.02 mA | Pass the calibration |
| 0MPa | 4.03 mA | Pass the calibration |
| … | … | … |
| 0MPa | 4.08 mA | Pass the calibration |
| 0MPa | 4.09 mA | Failure to pass the calibration |
| 8.75 MPa | 6.00 mA | Pass the calibration |
| 8.75 MPa | 6.01 mA | Pass the calibration |
| 8.75 MPa | 6.02 mA | Pass the calibration |
| 8.75 MPa | 6.03 mA | Pass the calibration |
| 8.75 MPa | 6.04 mA | Pass the calibration |
| 8.75 MPa | 6.04 mA | Pass the calibration |
| … | … | … |
| 8.75 MPa | 6.12 mA | Pass the calibration |
| 8.75 MPa | 6.13 mA | Failure to pass the calibration |
| … | … | … |

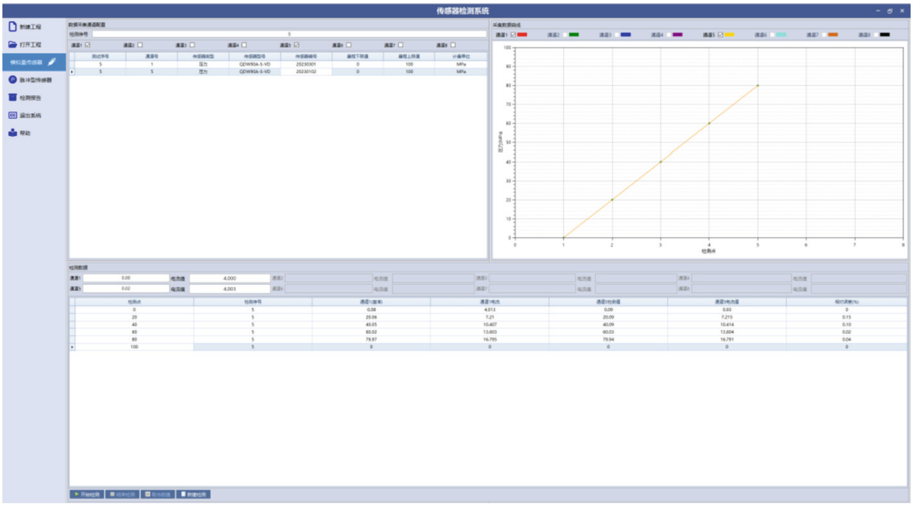
### 4.2 Experimental Verification and Data Visualization

Test the model, compare the experimental data with the predicted data, evaluate the reliability and accuracy of the model, and adjust and improve it accordingly.

Based on the experimental test results, present the data in the form of charts and analyze the sources and trends of errors to provide visual support for sensor calibration and testing.

As can be seen from the figure below, the data predicted by AI technology is in very good agreement with experimental data, with a maximum error of only 0.15%, thereby proving the effectiveness of this method.

Using the above technical solution, it is possible to utilize artificial intelligence technology for the calibration of logging tool sensors in order to improve testing accuracy and efficiency, reduce human error and testing costs, and provide strong support for the drilling engineering in the oil and gas industry.



数据采集通道配置

检测序号 5

通道1 ☑ 通道2 ☐ 通道3 ☐ 通道4 ☐ 通道5 ☑ 通道6 ☐ 通道7 ☐ 通道8 ☐

| 测试序号 | 通道号 | 传感器类型 | 传感器型号 | 传感器编号 | 量程下限值 | 量程上限值 | 计量单位 |
|---|---|---|---|---|---|---|---|
| 5 | 1 | 压力 | QDW90A-5-VD | 20230301 | 0 | 100 | MPa |
| 5 | 5 | 压力 | QDW90A-5-VD | 20230102 | 0 | 100 | MPa |

检测数据

通道1 0.00 电流值 4.000

通道5 0.02 电流值 4.003

| 检测点 | 检测序号 | 通道1(基准) | 通道1电流 | 通道5检测值 | 通道5电流值 | 相对误差(%) |
|---|---|---|---|---|---|---|
| 0 | 5 | 0.08 | 4.013 | 0.09 | 0.93 | 0 |
| 20 | 5 | 20.06 | 7.21 | 20.09 | 7.215 | 0.15 |
| 40 | 5 | 40.05 | 10.407 | 40.09 | 10.414 | 0.10 |
| 60 | 5 | 60.02 | 13.603 | 60.03 | 13.604 | 0.02 |
| 80 | 5 | 79.97 | 16.795 | 79.94 | 16.791 | 0.04 |
| 100 | 5 | 0 | 0 | 0 | 0 | 0 |

**Fig. 5.** Experimental Verification and Data Visualization

# 5 Conclusion

(1) The application of artificial intelligence in mud logging sensors can greatly improve the accuracy and reliability of the measurement results.

(2) The use of machine learning algorithms artificial neural networks (ANNs) can effectively address the problem of nonlinearity and complex interference in mud logging data.

(3) The calibration technology based on these algorithms has been successfully applied to real drilling engineering, achieving excellent results. The study improves the quality of mud logging data and provides a theoretical basis and practical guidance for the further promotion and development of intelligent mud logging technology.

**Acknowledgments.** The project is supported by China National Petroleum Corporation Research and Development Project "Research on Tracing and Transmission Technology of Oil and Gas Production Values" (Number 2021DJ2901).

# References

1. Balarabe, J.S., Abubakar, I.A., Nuhu, S.A., et al.: Artificial intelligence, sensors and vital health signs: a review. Appli. Sci. **12**(22) (2022)
2. Zhang, C., Cheng, K.: Accurate detection of intelligent running posture based on artificial intelligence sensor. J. Sensors (2022)
3. Chen, Y., Chen, Q.: Gymnastics action recognition and training posture analysis based on artificial intelligence sensor. J. Sensors (2022)
4. Li, K.: Tennis technology recognition and training attitude analysis based on artificial intelligence sensor. J. Sensors (2022)

5. Song, Z., Tian, C.: Influence of the athlete's training physical state test based on the principle of artificial intelligence sensor. Mobile Inform. Syst. (2022)
6. Michael, P., Douglas, B., Wayne, D., et al.: Artificial intelligence, sensors, robots, and transportation systems drive an innovative future for poultry broiler and breeder management. Animal Front. Rev. Mag. Animal Agricul. **12**(2) (2022)
7. Zeng, A., Yu, T., Song S., et al.: Multiview self-supervised deep learning for 6D pose estimation in the amazon picking challenge. In: 2017 IEEE International Conference on Robotics and Automation (CRAIEEE), pp. 386–383 (2016)
8. Zeng, A., Song, S., Yu, K.T., et al.: Robotic pick-and place of novel objects in clutter with multi affordance grasping and cross-domain image matching. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 1–8 (2018)
9. Ewerton M. Neumann G. Lioutikov Ral. Learning multiple collaborative tasks with amixture of interaction primitives C . EEE International Conference on Robotics & AutomationIEEE 2015. 1535–1542
10. Jingsha, Z., Yan, Z.: Research on automatic control of laser sensors based on artificial intelligence. Laser J. **43**(11), 199–203 (2022)
11. Hongwei, S., Na, L.: Automatic correction of ranging error of laser displacement sensors using artificial intelligence technology. Laser J. **42**(10), 167–170 (2021)
12. Xuetong, R.: Research on sensor technology based on artificial intelligence. Mod. Indust. Econ Inform. **10**(05), 60–61 (2020)
13. Zhiwu, W.: Fault diagnosis technology of sensors based on artificial intelligence methods. Rocket Propulsion **05**, 59–62 (2005)
14. Beizhan, P., Lin Dejie, O., Jincheng.: Application of artificial intelligence in the field of sensors. Sensor Technology **03**, 5–7 (2002)
15. Yan, S., Lei, H., Yan, R.: Design of an automatic calibration system for temperature sensors based on robots. Electronic Measure. Technol. **44**(09), 56–65 (2021)
16. Xianghua, H., Feng, J., Shuiwang, Y., et al.: Application of artificial intelligence in field dynamic calibration of vector thrust. Aerospace Measurem, Technol. **39**(03), 51–57 (2019)

# Sensitivity Analysis of Influencing Factors of Production for Fractured Horizontal Wells in Shale Reservoir

Wei Liu[1,2](✉), Xiao-peng Cao[1], Zi-yan Cheng[1], and Yan Liu[1]

[1] Exploration and Development Research Institute, Sinopec Shengli Oilfield Company,
Dongying 257099, Shandong, China
lwsg93@126.com

[2] Postdoctoral Scientific Research Working Station, Sinopec Shengli Oilfield Company,
Dongying 257099, Shandong, China

**Abstract.** Major productivity breakthroughs have been achieved in key production layers of Jiyang shale, such as lower Es3 and upper Es4 producing layer of Shahejie Formation, and remarkable development have been gained. However, it is also limited by short production time, large production difference of single well, and unclear of production influencing factors. Comprehensive analysis of the main controlling factors of production for horizontal shale oil wells has become the research focuses. Field production data was taken to clarify the influence of various factors on the production of horizontal wells. Grep correlation analysis and principal component analysis were used to quantitatively analyze the sensitivity of 90-day average oil production, 180-day average oil production, and 270-day average oil production to the influencing factors, such as the amount of fluid used and the sand added. Research indicates that the amount of fluid used, the amount of sand added and the number of fracture events are the main engineering parameters affecting the production, while the content of gray matter, TOC and shale porosity are the main geological parameters affecting the production. The influence of geological factors on production gradually increase, while that of engineering factors on the production is gradually weakened in the late flowing production stag. The main controlling factors and variation rules of the production are preliminarily identified, which could provide guidance for the deployment of shale oil wells and fracturing design.

**Keywords:** Shale oil and gas · horizontal well production · influencing factors analysis · grey correlation theory · principal component analysis

## 1 Introduction

With the increasing energy demand, shale oil has gradually become a hot spot in oil and gas exploration and development. China's shale oil resources are abundant and have a broad prospect for exploration and development. As a typical Cenozoic oil-rich terrestrial fault basin in eastern China, the Jiyang Depression has a high potential and abundance of

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 124–134, 2024.
https://doi.org/10.1007/978-981-97-0272-5_10

shale oil resources, which has a high extraction value [1]. Compared with shale oil from other regions at home and abroad, Jiyang shale oil is very different in terms of formation environment, shale characteristics, degree of evolution and shale oil properties. It is a unique type of shale oil with strong non-homogeneity and large differences in oil content. The exploration and development of Jiyang shale oil has gone through three stages, namely exploration incidental, active exploration and innovative breakthrough, and has achieved significant capacity breakthroughs in several layers and types. The peak oil volume of many shale oil wells that have been put into production exceeds 100 tons, and the cumulative oil in six months exceeds 10,000 tons, demonstrating the good prospect of Jiyang shale oil exploration and development [2].

In the development process of Jiyang Shale Oil using horizontal well volume fracturing, it is affected by multiple factors of reservoir characteristics, fracture modification and other geological engineering, and there are many factors affecting production and the situation is complicated, which shows the problem of large variation of single well production and the main control factor of production capacity is still unclear in production. In this regard, there is an urgent need to carry out research on the analysis methods of the main control factors of the post-pressure production of Jiyang shale oil horizontal wells, clarify the influence of each factor on the production, and optimize the development technology countermeasures to guide the development of Jiyang shale oil in a cost-effective way [3].

For the analysis of factors affecting production, scholars at home and abroad mainly use multiple linear regression [4, 5], neural networks [6, 7] and other methods to quantitatively analyze the influence of reservoir quality parameters, engineering quality and other parameters of shale oil on production [8–13]. Luo [6] et al. used artificial neural network to analyze the influence of geological feature parameters and completion parameters on the cumulative production in the first year, which provided important guidance for the development of reasonable fracturing construction parameters in the study block. Guo Jiancheng[7]et al. analyzed the factors influencing the flowback rate and production capacity of shale gas wells in the Longmaxi Formation based on a neural network approach, and established a capacity prediction plat for the block to guide production prediction and decision making. Kim [8] conducted a quantitative analysis of reservoir quality and completion quality to quantify the influence of each factor on production. Zhang et al. [40] analyzed the main factors affecting production by establishing a production prediction model for horizontal shale gas wells. Wu Linhong et al. [13] established a single-well capacity model for fractured horizontal wells by considering geological factors and fracturing parameters, and analyzed the influence weight of each factor on production by comparing the defined influence factors. Due to the strong non-homogeneity of domestic onshore phase, the results of shale oil production analysis in different fields are somewhat different [14–17]. The results of the current study are not all applicable to Jiyang shale oil because of its diverse types and low maturity.

In order to clarify the main control factors of Jiyang shale oil production, the correlation between initial production capacity and shale physical parameters, fracturing parameters during shale oil development is studied from the perspective of geological engineering integration using gray correlation analysis, and the degree of influence of each factor on initial production capacity is clarified. To verify the reliability of the method, principal component analysis was used to further validate the results of the analysis of production master control factors. At the same time, the influence of each factor on the production at different production stages was quantitatively analyzed to reveal the main controlling factors at different production stages, so as to provide support for the formulation of reasonable development plan and effective development of shale oil [18].

# 2 Analysis Method of Production Influencing Factors

## 2.1 Grey Relation Analysis

Grey correlation analysis is one of the main contents of grey correlation system theory, which has been widely used in reservoir evaluation in particular [3]. Grey correlation analysis can be used to quantitatively analyze the degree of correlation between output and various influencing factors, so as to realize the screening of main controlling factors of productivity [19, 20]. The specific process includes selecting reference sequence and comparison sequence, then calculating correlation coefficient and correlation degree, and finally determining the main influencing factors of productivity.

### 2.1.1 Determine the Reference and Comparison Sequences

This paper focuses on analyzing the influence of shale oil geological parameters, fracturing parameters and other factors on production through gray correlation. In this regard, production is determined as the reference sequence $X_0(k)(k = 1, 2, 3, \cdots, n)$, production influencing factors as a comparative series $X_i(k)(i = 1, 2, 3, \cdots, m)$, $n$ indicates the number of elements in each series, $m$ denotes the number of influencing factors, $I$ indicates the serial number of the influencing factor. The production data, geological and engineering data of Jiyang shale oil Wells are collected and the analysis matrix in the following form is constructed:

$$(X_0, X_1, \ldots, X_m) = \begin{pmatrix} X_0(1) & X_1(1) & \cdots & X_m(1) \\ X_0(2) & X_1(2) & \cdots & X_m(2) \\ \vdots & \vdots & \vdots & \vdots \\ X_0(n) & X_1(n) & \cdots & X_m(n) \end{pmatrix} \tag{1}$$

### 2.1.2 Dimensionless Processing of Data

Considering that the difference in order of magnitude between the original series will affect the calculation results of correlation degree, dimensionless processing should be carried out for these original series. The specific calculation formula is as follows::

$$X_i\prime(k) = \frac{X_i(k) - \min(X_i)}{\max(X_i) - \min(X_i)} \tag{2}$$

where, $X_i\prime(k)$ is the $k$th parameter value of the $i$th influencing factor, $\max(X_i)$ is the maximum value of the $i$ sequence of influencing factors $X_i$, $\min(X_i)$ is the minimum value of the $i$ sequence of influencing factors $X_i$.

### 2.1.3 Calculation of Correlation Coefficient

According to the definition of correlation coefficient (3), the correlation coefficient between influence factor series and production reference series is calculated separately.

$$\sigma_{0i}(k) = \frac{\min\limits_{i=1}^{m}\min\limits_{k=1}^{n}|x_i(k) - x_0(k)| + \rho \times \max\limits_{i=1}^{m}\max\limits_{k=1}^{n}|x_i(k) - x_0(k)|}{|x_i(k) - x_0(k)| + \rho \times \max\limits_{i=1}^{m}\max\limits_{k=1}^{n}|x_i(k) - x_0(k)|} \tag{3}$$

where, $\rho$ indicates the discrimination factor, the range of values is [0, 1], $\sigma_{0i}(k)$ denotes the number of correlation coefficients between the $i$th influencing factor and oil production at the $k$th value.

### 2.1.4 Calculation of Correlation Degree

The correlation between production and influence factors is calculated by averaging the correlation coefficients between each sequence of influence factors and the corresponding elements of the production reference sequence, and calculating the correlation between prodcuiton and influence factors sequentially according to Eq. (4).

$$r_{0i} = \frac{1}{n}\sum_{k=1}^{n}\sigma_{0i}(k), \ (i = 1, 2, \ldots, m) \tag{4}$$

where, $r_{0i}$ denotes the correlation degree between the $i$th influencing factor and the production, the range of values is [0,1].

In order to screen out the main control factors affecting the production, on the basis of obtaining the correlation degree between production and variables, the weights of each influencing factor are calculated based on Eq. (5). The factors are ranked according to the weights so as to determine the main control factors of production capacity.

$$W_i = \frac{r_{oi}}{\sum\limits_{i=1}^{m} r_{oi}} \times 100 \tag{5}$$

where, $W_i$ indicates the weight of the ith influencing factor.

## 2.2 Principal Component Analysis

Principal Component Analysis (PCA) is a common data analysis method [21, 22], which transform the original data into a set of linearly independent representations by linear transformation. And it can be used to extract the main feature components of the data. The matrix of observed variables consisting of m influencing factors is expressed as $X = (X_1, X_2, \ldots, X_m)$ (see Eq. (6). The original influencing factor variables are regrouped into a new set of mutually unrelated composite variables to replace the original variables through principal component analysis. The first linear combination selected, i.e., the first composite variable, is denoted as.

If the first principal component is not sufficient to represent the information of the original m variables, then consider selecting F2. In order to effectively reflect the original information, the information already in F1 need not appear in F2 again, i.e., Cov(F1,F2) = 0. By analogy, the third, fourth and mth principal components can be constructed, which are calculated by the formula in Eq. (7).

$$X = \begin{pmatrix} X_1(1) & X_2(1) & \cdots & X_m(1) \\ X_1(2) & X_2(2) & \cdots & X_m(2) \\ \vdots & \vdots & \vdots & \vdots \\ X_1(n) & X_2(n) & \cdots & X_m(n) \end{pmatrix} = (X_1, X_2, \ldots, X_m) \tag{6}$$

$$F_j = a_{j1}X_1 + a_{j2}X2 + \cdots + a_{jm}X_m, \quad j = 1, 2, \cdots, m \tag{7}$$

where, $a_{ij}$ indicates the principal component coefficients, and the new variables and coefficients for the principal component analysis are required to satisfy the following conditions.

a. $F_i$ , $F_j$ unrelated to each other $(i \neq j; i, j = 1, 2, \cdots, p)$.

b. The variance of F1is greater than the variance of F2, and so on.

c.$a_{j1}^2 + a_{j2}^2 + \cdots + a_{jm}^2 = 1$.

# 3 Applications

Jiyang shale oil is mainly developed through multi-stage fractured horizontal wells, and the production generally shows a fast decreasing in the initial stage and a slow decreasing in the later stage. However, the production of shale oil wells varies greatly from depression to depression, and the production variation pattern varies among wells in the same depression. In order to reveal the main control factors for high production of Jiyang shale oil, the influence of geological and engineering factors on shale oil production was quantitatively analyzed. A total of 17 specific geological and engineering factors were considered, as shown in Table 1..

**Table 1.** Statistical of production influencing factors

| Influencing factors | Parameters |
|---|---|
| Geological factors | TOC, S1, GR, AC, CNL, DEN, ash content, clay content, sand content |
| Engineering factors | Fracturing fluid volume, sand added, CO2 volume, breaking pressure, section length, stages, sand ratio, number of microseismic events |

## 3.1 Initial Capacity Gray Relation Analysis

The geological, engineering and production data of the horizontal wells after fracturing and were selected as the basis for analysis of production impact factors. The initial production (90-day average daily oil production) of each fracturing section is used as the reference sequence$X0$, the remaining 17 influencing factors are used as reference sequences $X_i(i = 1,2,\ldots, 17)$. The correlation between the initial production and influencing factors were calculated according to Eq. (4), and the results are shown in Tables 2. and 3.. In order to clarify the primary and secondary relationships among the influencing factors, the weights $W_i(i = 1,2,\ldots, 18)$ of each influencing factor are further calculated based on Eq. (5), as show in Fig. 1. It can be seen that the weight of fluid used and sand addition is the largest among the factors, and the weight of TOC, AC and ash content among the geological factors are all above 5.0. It indicates that fracturing design is the main factor affecting the initial production, and oil content, pore space and compressibility are the main geological factors affecting the initial production.

**Table 2.** Correlation between geological factors and initial production

| Influencing factors | TOC | AC | GR | AC | CNL | DEN | Ash content | Mud content | Sand content |
|---|---|---|---|---|---|---|---|---|---|
| Correaltion degree | 0.073 | 0.032 | 0.006 | 0.064 | 0.009 | 0.011 | 0.068 | 0.017 | 0.044 |

**Table 3.** Correlation between engineering factors and initial production

| Influencing factors | Fracturing fluid volume | Sand added | $CO_2$ volume | Breaking pressure | Section length | Stages | Sand ratio | number of microseismic events |
|---|---|---|---|---|---|---|---|---|
| Correaltion degree | 0.134 | 0.331 | 0.006 | 0.058 | 0.025 | 0.005 | 0.105 | 0.121 |

To verify the reliability of the results of gray relation analysis, principal component analysis was used to analyze the geological and engineering factors separately, and the importance of the influencing factors was evaluated according to the coefficients of each feature in the principal components. Firstly, the principal component analysis was done for the geological factors, and the analysis results are shown in Table 4. From the analysis results, we can see that the cumulative contribution of the variance of the first 5 principal components is 94%, that is, the first 5 principal components can explain 94% of the information of the 10 factors, so only the first 5 principal components are needed to replace all the data information. The coefficients of each influence factor in these 5 principal components are shown in Table 5, and the larger the absolute value of the coefficients, the more information the principal component reflects about that influence factor. From Table 5, it can be seen that the coefficients of TOC, AC, DEN and ash content are relatively large and are the main geological parameters affecting yield, which basically coincide with the results of gray correlation analysis.

The results of the principal component analysis of the engineering factors are shown in Table 6. The first 4 principal components can represent 90% of the information of all

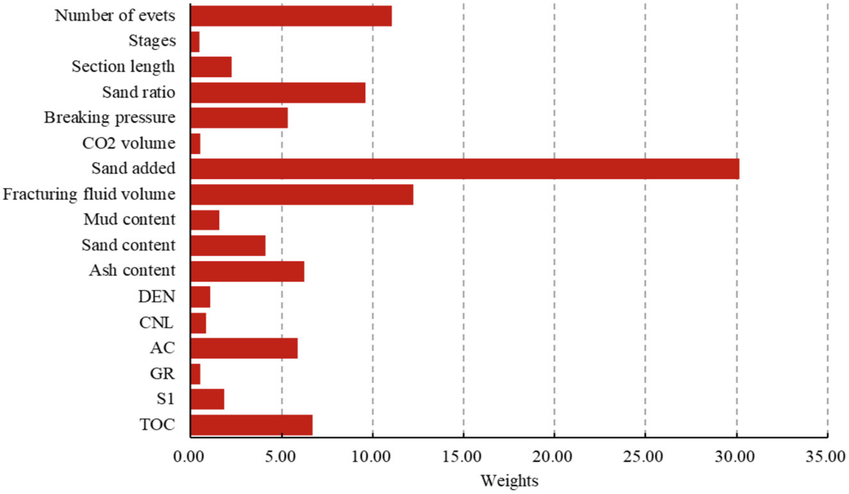


**Fig. 1.** Bar chart of weight of production influencing factors

**Table 4.** Principal component contribution of geological factors

| Index | Component 1 | Component 2 | Component 3 | Component 4 | Component 5 |
|---|---|---|---|---|---|
| Standard deviation | 0.47 | 0.27 | 0.07 | 0.05 | 0.04 |
| Variance contribution | 0.49 | 0.28 | 0.08 | 0.05 | 0.04 |
| Cumulative variance contribution | 0.49 | 0.77 | 0.85 | 0.90 | 0.94 |

8 engineering factors. The coefficients of each influencing factor in the first 4 principal components are shown in Table 7, and it can be seen that the coefficients of fluid volume, sand addition, sand ratio and microseismic events are relatively large, which are basically consistent with the results of gray correlation analysis The results of the analysis of production influencing factors based on gray correlation are more reliable and can be used for the analysis of the main control factors of shale oil initial production.

### 3.2 Analysis of Production Influencing Factors at Different Production Stages

In order to reveal the degree of influence of geological and engineering factors on production at different production stages, the average daily oil production data of horizontal wells for 180 days and 270 days were used as the reference sequence $X0$. Based on gray relation analysis, the correlation degree between the comparative sequence of influencing factors and the reference sequence was calculated, and the weights of each influencing factor when the reference sequence was the average daily oil production of 90 days,

**Table 5.** Distribution of principal component coefficients of geological factors

| Influencing factors | Component 1 | Component 2 | Component 3 | Component 4 | Component 5 |
|---|---|---|---|---|---|
| TOC | 0.64 | −0.09 | −0.05 | 0.19 | −0.26 |
| S1 | 0.01 | −0.18 | −0.36 | −0.17 | −0.23 |
| GR | −0.04 | 0.12 | 0.34 | −0.37 | 0.02 |
| AC | −0.41 | −0.16 | −0.20 | −0.14 | 0.41 |
| CNL | −0.10 | 0.36 | 0.13 | −0.20 | -0.27 |
| DEN | -0.08 | 0.12 | -0.03 | 0.57 | 0.23 |
| Ash content | 0.20 | −0.27 | 0.61 | 0.17 | -0.07 |
| Mud content | −0.24 | 0.32 | 0.25 | 0.14 | 0.03 |
| Sand content | 0.00 | 0.01 | 0.26 | -0.46 | 0.07 |

**Table 6.** Principal component contribution of engineering factors

| Index | Component 1 | Component 2 | Component 3 | Component 4 |
|---|---|---|---|---|
| Standard deviation | 0.08 | 0.04 | 0.03 | 0.01 |
| Variance contribution | 0.43 | 0.24 | 0.18 | 0.05 |
| Cumulative variance contribution | 0.43 | 0.67 | 0.84 | 0.90 |

**Table 7.** Distribution of principal component coefficients of engineering factors

| Influencing factors | Component 1 | Component 2 | Component 3 | Component 4 |
|---|---|---|---|---|
| Fracturing fluid volume | 0.92 | 0.78 | 0.03 | 0.03 |
| Sand added | 0.34 | 0.38 | 0.94 | 0.17 |
| CO2 volume | -0.26 | 0.17 | 0.03 | 0.10 |
| Breaking pressure | 0.45 | 0.16 | 0.28 | -0.25 |
| Sand ratio | 0.08 | 0.01 | 0.03 | 0.20 |
| Section length | 0.05 | 0.36 | 0.17 | 0.05 |
| Stages | 0.63 | 0.75 | 0.09 | 0.12 |
| Number of evets | 0.32 | 0.52 | 0.30 | 0.81 |

180 days and 270 days respectively were counted, and the summary results are shown in Table 8. From the analysis results, it can be seen that the influence of TOC, S1 and

AC in geological factors on the 270-day average daily oil production is increased compared with the 90-day and 180-day average daily oil production, and the weight of oil content can reach 18.54. The influence of engineering factors of fluid usage and sand addition on production gradually decreases with the extension of time. This is because the production at the early stage of production mainly comes from the fracture fracture network. Due to the strong stress sensitivity of shale oil, the formation pressure gradually decreases and the fracture conductivity gradually becomes worse as the subsequent production proceeds, resulting in the influence of engineering factors on production gradually weakening. As the pressure propagates from the vicinity of the wellbore to the periphery, the reservoir matrix part gradually participates in the mobilization, and the influence of geological factors on production gradually increases, which verifies with the results of the analysis of production influence factors in different production stages.

**Table 8.** Weighting of factors influencing production at different production stages

| Influencing factors | 90-day average daily oil production | 180-day average daily oil production | 270-day average daily oil production |
|---|---|---|---|
| TOC | 6.65 | 8.96 | 11.56 |
| S1 | 1.82 | 4.65 | 8.06 |
| GR | 0.52 | 0.66 | 1.05 |
| AC | 5.83 | 7.63 | 9.89 |
| CNL | 0.82 | 1.24 | 1.93 |
| DEN | 1.05 | 2.07 | 1.75 |
| Ash content | 6.22 | 5.97 | 6.04 |
| Sand content | 1.59 | 1.74 | 2.80 |
| Mud content | 4.05 | 3.24 | 3.59 |
| Fracturing fluid volume | 12.20 | 12.03 | 9.81 |
| Sand added | 30.11 | 24.98 | 22.15 |
| CO2 volume | 0.53 | 0.50 | 1.14 |
| Breaking pressure | 5.30 | 5.14 | 3.77 |
| Sand ratio | 2.24 | 1.99 | 2.71 |
| Section length | 0.49 | 0.76 | 1.05 |
| Stages | 9.56 | 7.63 | 7.53 |
| Number of evets | 11.02 | 10.79 | 5.17 |

## 4 Conclusion

(1) The results of the main control factors of the initial production based on gray relation analysis show that the initial production of volumetric fractured horizontal wells is better correlated with engineering factors, and the fracturing design is the main factor affecting the initial production.

(2) Analysis of the factors influencing production at different stages shows that the influence of engineering factors on production gradually diminishes as production proceeds, the influence of geological factors on production gradually increases, and the oil in matrix gradually becomes involved in mobilization.

(3) According to the analysis of the production controlling factors of Jiyang shale oil, the sweet spot with better oil content and porosity should be preferentially selected for development, and the production of shale oil wells could be further improved by optimizing the amount of fluid used and sand addition.

(4) In future research, it is necessary to further analyze the synergistic effects of influencing factors on production to effectively guide the deployment of horizontal shale oil wells and fracturing design.

**Acknowledgments.** The project is supported by Shandong Province Postdoctoral Innovation Talent Support Program (Data-driven based shale oil production forecasting and optimization).

## References

1. Song, M.: Ractice and current status of shale oil exploration in Jiyang depression. Petrol. Geology Recovery Effici. **26**(01), 1–12 (2019)
2. Liu, H.: Exploration practice and prospect of shale oil in Jiyang depression. China Pet. Explorat. **27**(01), 73–87 (2022)
3. Sun, H.: Exploration practice and cognitions of shale oil in Jiyang depression. China Pet. Explor. **22**(4), 14 (2017)
4. Tao, L., Yuwen, C., Xiaofei, G., et al.: Influence factors of single well's productivity in the Bakken tight oil reservoir. Pet. Explor. Dev. **40**(3), 357–362 (2013)
5. Xinhua, M., Xizhe, L., Feng, L., et al.: Dominating factors on well productivity and development strategies optimization in Weiyuan shale gas play, Sichuan Basin, SW China. Pet. Explor. Dev. **47**(3), 594–602 (2020)
6. Liang, G., Chang, Y., Guo, X.: Production-strategy insights using machine learning: application for Bakken shale. SPE Reservoir Eval. Eng. **22**(3), 800–816 (2019)
7. Guo, J., Lin, B., Xiang, J., et al.: Study of factors affecting the flowback ratio and productive capacity of Longmaxi Formation shale in the Sichuan basin after fracturing. Petrol. Sci. Bull. **4**(3), 273–287 (2019)
8. Kim, G., Lee, H., Chen, Z., et al.: Effect of reservoir characteristics on the productivity and production forecasting of the Montney shale gas in Canada. J. Petrol. Sci. Eng. **182**(1), 106276 (2019)
9. Zhang, Y., Wang, C., Yang, L., et al.: Analysis and prediction method of influencing factors on productivity of tight oil horizontal well in Daqing oilfield. Logging Eng. **29**(03), 55–59 (2018)

10. Liu, W., Liu, W., Gu, J.: Oil production prediction based on a machine learning method. Oil Drilling Product. Technol. **42**(01), 70–75 (2020)
11. Xu, J., Yang, L., Ding, Y., et al.: Influencing factors on the productivity of the volume-fractured horizontal well in the tight oil reservoir. Pet. Geol. Oilfield Developm. Daqing **39**(01), 162–168 (2020)
12. Ma, W., Li, Z., Gao, C., et al.: "Pearson- MIC" analysis method for the initial production key controlling factors of shale gas wells. China Sciencepaper **13**(15), 1765–1771 (2018)
13. Wu, L., Guo, X., Luo, W., et al.: Influence factors controlling the productivity of horizontal well by volume fracturing in tight oil wells. Unconventional Oil Gas **5**(03), 56–62 (2018)
14. Ji, L., Xiao, J.: Application of random forest algorithm in the multistage fracturing stimulation of shale gas field. Petroleum Geol, Oilfield Developm, Daqing **39**(6), 168–174 (2020)
15. Sun, J., Liu, D., Zhang, L., et al.: Grey correlation analysis of factors affecting decline of low permeability reservoir. Special OilGas Res. **19**(2), 90–93 (2012)
16. Wei, J., Zhang, Y., Shang, J., et al.: Principal factor analysis on initial productivity in shale oil development: A case study of Block Li-151 in Changqing Oilfield. Reservoir evaluation and development **11**(4), 550–558 (2021)
17. Hao, B., Ma, J., Ye, B., et al.: Analysis on Influencing factors of tight oil horizontal well production in Ordos Basin. Well Tesing **26**(04), 16–18 (2017)
18. Wang, J., Li, J., Chen, X., et al.: Research and practice of integrated fracturing design technology for 3D well pattern of permian lucaogou formation in jimsar depression in Junggar basin. Petroleum Sci. Technol. Forum **41**(02), 62–68 (2022)
19. Huang, F.: Sun Dongsheng application of grey correlation analysis in screening of hard-to-produce potential blocks. Comp. Hydrocar. Reserv. **15**(02), 67–71 (2022)
20. Liu, X.: Comparing for grey forecast and forecast of one element linear regression. J. Univ. Sci. Eng. (Nat. Sci. Edn.) **22**(01), 107–109 (2009)
21. Karamizadeh, S., Abdullah, S.M., Manaf, A.A., et al.: An overview of principal component analysis. J. Signal Inform. Process. **4**(3B), 173 (2013)
22. Shan, C., He, J., Zhou, T., et al.: A Study on the optimization of fracturing operation parameters based on PCA-CNN. J. Southwest Petrol. Univ. (Sci. Technol. Edn.) **42**(6), 56–62 (2020)

# Research and Practice of Digital Three-Phase Flowmeter for Complex Oil and Gas Occasion

Bing Chen, Miao Liu(✉), Ya-nan Zhang, Hong-zhi Han, and Xin-dong Guo

Kunlun Difital Technology Co. Ltd., Beijing 102206, China
Liumiao01@cnpc.com.cn

**Abstract.** In the development of oil and gas fields, the accurate measurement of oil, gas and water production rate is the basis for calculating the key parameters such as water cut (WCT) and gas-oil ratio (GOR), as well as the important basis for formulating the stimulation measures and field development plans. The traditional surface flowing test is mainly carried out with the help of three-phase Separator or multiple phase flowmeters (MPFM), which is not only complicated operation process and long operating duration time, but also cannot be implemented for environmental protection reasons in some special location. In recent years, with the acceleration of digital oilfield transformation, many wellhead digital multi-phase flowmeters have appeared in the market. However, due to the interference of flow rate, high water cut, high gas content and other factors, the measurement accuracy of most three-phase flowmeters in the market is not good enough, which cannot meet the actual requirement of the customers. In view of the above problems, the design and development of a new on-line three-phase flowmeter is carried out, and the hardware and software system of the flowmeter is upgraded with an iterative and innovative method. Through large-scale field pilot tests to evaluate the performance of equipment, find out the problems during the testing process, and continuously improve the hardware design, software function and core model of the product, so that the instrument can detect the gas and liquid production rate online in real time, and realize real-time data collection and communication through the deep integration with the Internet of Things technology. Present the data to the user in a visual manner in the same time, thus, the whole field data sharing can be realized. The product was put on line in a domestic oilfield after pilot test. The application results show that the product has high stability and the ability

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 135–143, 2024.
https://doi.org/10.1007/978-981-97-0272-5_11

to work under complex conditions, and can meet the needs of flow measurement under high GOR and high water cut in the oilfield, and has the feasibility to widely expand the application in domestic and foreign oil fields.

**Keywords:** three-phase flowmeter · online detection · The Internet of things · iterative innovation · water content rate · error rate

## 1 Introduction

During oil and gas development, continuous monitoring of wells to obtain pressure, temperature and production data is key to grasping well production status and analyzing reservoir dynamics parameters. Among them, it is particularly important to accurately measure the three-phase flow, which allows calculation of the well's water cut and gas-to-oil ratio. These parameters are a significant basis for formulating oil well stimulation measures and development adjustment plans.

Oil, gas and water are multi-phase flows with complex flow patterns in the well, which increases the difficulty of flow measurement. Normally, truck-mounted three-phase separators are typically used at the wellhead to separate and measure oil, gas and water rate. In addition, stationary three-phase separators at oil and gas processing stations can also be used to separate mixed phases and obtain information such as water cut. However, these processes are laborious, expensive, and do not allow continuous metering. There is an urgent need for reliable online three-phase metering devices for oilfield development and production to reduce investment and improve metering efficiency.

By deploying three-phase flow meters in oil fields, it can realize online automatic collection of oil, gas and water production data from single wells and clusters, optimize the surface process of oil and gas field enterprises, reduce operation costs, improve the accuracy and timeliness of production management and geological reservoir analysis, and provide support for the digital transformation [1–4].

However, due to the oil, gas and water in different flow rates (liquid phase, gas phase flow rate) and gas-liquid ratio to form a variety of flow patterns, the use of three-phase flowmeter for multiphase flow testing is relatively complex [5, 6]. For cross-tube, including laminar flow, bubble flow, segment plug flow, fog flow, etc., by many factors and measurement difficulties, the measurement accuracy of three-phase flowmeter compared with the traditional three-phase separator has a large error, which as a technical bottleneck to limit the three-phase flowmeter large area applications [7, 8]. It is usually conducts continuous field pilot tests on three-phase flowmeters, using iterative innovation to continuously improve the measurement results, so that three-phase flowmeters can more accurately obtain the flow of oil, gas and water to meet the requirements of oilfield production management and reservoir dynamic analysis.

# 2 Online Three-Phase Flowmeter Development Process and Problems

## 2.1 The Test Platform and Development Process of the First Generation Equipment

The online three-phase flowmeter in this study is developed in cooperation with the top multi-phase flow laboratory in China according to the actual needs of the oilfield, breaking the foreign technological monopoly [9–12] and achieving the completely independent research, achieving the scientific goal of domestic replacement of all similar products in the oilfield.

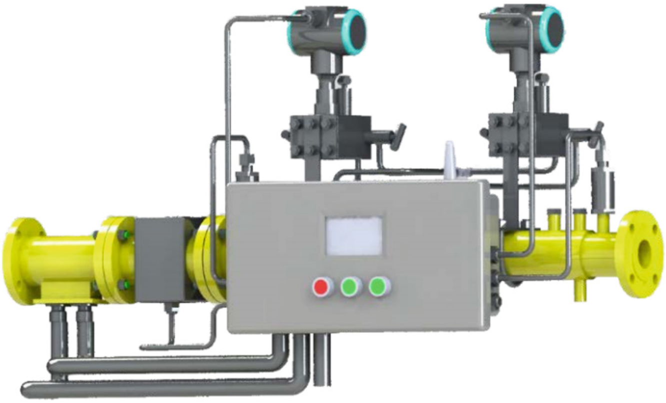


**Fig. 1.** Online three-phase flowmeter appearance structure diagram.

At the end of 2017, the first generation of online three-phase flowmeter product prototype was successfully developed (as shown in Fig. 1). The product, based on a compact integrated measurement solution with multiple sensors, has achieved complete independent development and localization in software and hardware technology. With many advantages such as green and radiation-free measurement process, safety, reliability, high accuracy, modular design and low maintenance cost, it fills the gap of low-cost, non-separated three-phase flow online measurement technology between China and west developed countries in the world. In 2018, the product was completed testing and put on line for the first field pilot test.

After the successful of the maiden voyage, five domestic oilfields were selected for phase II testing at the same time. By deploying 100 sets of three-phase flowmeter equipment in different scenarios such as single well wellhead, inter-meter backdown metering and cluster wells metering, the production of oil, gas and water is automatically collected online without separation, meanwhile, the accuracy of data meets the demand of oilfield.

### 2.2 Technology Principle

The new online three-phase flowmeter consists of four main modules: venturi measurement module, microwave detection module, electro-laminar imaging module, data processing and communication module.

(1) Venturi measurement module

Dual differential pressure venturi flow meters are used to measure the total flow rate in the gas and liquid phases. The differential pressure flowmeter module consists of a venturi with a differential pressure sensor, a pressure sensor and a temperature sensor, connected to a data acquisition and processing module.

(2) Microwave moisture content detection module

The microwave water content module includes a number of microwave sensors with different spatial position directions and angles, which consist of a transmission line set inside a seal, a seal set inside an insulating medium, and an insulating medium set inside a tube. The total circuit module uses the average value of microwave phases to determine the water phase content in the three phases.

By combining the power attenuation and phase angle shift of the detected microwaves in the fluid with a water cut calculation model, the water cut can be calculated as per detected data. This method is more accurate and less influenced by the mineralized content than traditional sensors with RF conductivity technology.

(3) Electromagnetic measurement imaging and display module

Based on the electro-layer imaging technique, the distribution is obtained by applying electrical excitation to the mixed-phase fluid, detecting the boundary value changes, and using mathematical means to invert the distribution of the electrical characteristic parameters inside the mixed-phase fluid. The display module is connected to the data acquisition and processing module and is used to display the results of flow calculations from the output data acquisition and processing module.

(4) Data processing analysis and communication module.

The data processing module is used to receive the differential pressure signal of homogeneous flow output by differential pressure flowmeter, and substitute it and volume flow rate into multi-phase flow empirical model to calculate the average density of homogeneous flow. At the same time, the working gas density can be calculated from the pressure and temperature signals collected by the differential pressure flowmeter and the gas component of the oil-gas-water three-phase flow.

In addition, The data processing module obtains the three-phase flow rates of oil, gas and water by solving a linear system of equations for the interrelationship between the average density, average dielectric constant and phase content of the homogeneous flow, and uploads them to the system through the communication system for use by all relevant departments in the oilfield for queries.

## 3 Performance Enhancement and Product Upgrade

By summarizing the technical problems of the first-generation machine, the researchers conducted countermeasure research in different test bases with the idea of iterative innovation. Through continuous improvement of the hardware design, software functions and core model of the product, the second generation of the product was formed, which has been greatly improved in terms of environmental applicability, working stability and performance index.

**Table 1.** Iterative innovation upgrade content of three-phase flowmeter

| Items | Improvement content | Improvement effect |
|---|---|---|
| Hardware Design | Improved differential pressure transmitter, using a new capillary differential pressure transmitter to replace the traditional pilot pressure differential pressure transmitter | Improved the stability of equipment working for a long time and the consistency between different equipment and different time periods of the same equipment.<br>Improved the applicability to frequently changing well conditions and the feasibility of sensor data, and enhances the robustness of the model.<br>Avoided the problem of signal distortion caused by occasional lead pipe blockage |
| | Improved environmental suitability in terms of equipment differential pressure sensors, interface piping design, etc | The differential pressure transmitter adopts professional capillary pressure to replace the original metal pipe pressure, which improves the adaptability to low temperature.<br>Improved interface and piping design to further reduce the requirement for insulation work |
| | Improved hardware design to facilitate on-site maintenance | Reduced device size for easy transportation, installation, post-operation and maintenance.<br>Improved display to support instant wake-up and reduce energy consumption |
| | .Improved enclosure design, added keys and enhanced site protection | Improved the box design to reduce the difficulty of opening the box for maintenance by reducing the size.<br>Added buttons to facilitate user's on-site operation of inverted wells and data inquiry.<br>Improved the effect of rainproof and dustproof |

(*continued*)

**Table 1.** (*continued*)

| Items | Improvement content | Improvement effect |
|---|---|---|
| Software Features | Support inverted well operation, real-time data query, and historical data query | The device added buttons to support user field operation and reverse well change, and also reserves interfaces to support linkage with automatic well change systems.<br>Supported real-time data and historical data query, including liquid, oil, gas, water, water content, gas-oil ratio by day |
| Core Model | Introduce a new algorithm architecture of artificial intelligence + classical model | Artificial intelligence-based deep learning for real-time classification of flow patterns and flow regimes, and data computation based on classical fluid dynamics models.<br>Fully reuse the existing calibration data of each well to support data migration and assist in AI model classification training.<br>On the basis of achieving consistency among equipment, unify models across ranges and equipment areas to improve the applicability of equipment to new well conditions and significantly reduce error levels |

The second-generation prototype was quickly entered several oilfield plays for field testing and evaluation, targeting single-well metering and covering a wide range of extraction methods, including natural flowing well, rod pump and ESP wells. Test well conditions include high GOR and high water cut wells located in the Northwest field with flow rates from 100 to 350 bbls/d, gas production rates of 0 to 15,000 m3 /day and water cut from 50% to 99%; and low production and high water cut wells located in the Northeast field with flow rates of 50 to 100 bbls/d. Through testing to verify the applicability of the three-phase flowmeter in multiple production ranges, the product supports the provision of second-level online flow data, which can help users more effectively determine the trends and changes in oil and gas well production, providing a reliable basis for oilfield production management.

## 4 Field Applications

After the new online three-phase flowmeter product was put on line, a pilot test was conducted in a domestic E&P company of China National Petroleum Corporation, which mainly included the following four purposes.

(1) Verify the performance of real-time online three-phase flowmeter in oilfield environment.
(2) Obtain data on the applicable working condition range and metering accuracy of the real-time online three-phase flowmeter.
(3) Verify the stability of real-time online three-phase flowmeter under extreme environmental conditions.
(4) Update and iterate on the equipment to meet the needs of each oilfield technology promotion in response to the product usage effect.

Fifteen units are deployed for block metering and single well metering in the X field owned by the exploration company, involving various well types such as pumping and electric submersible pump wells, with liquid volumes ranging from 100–1500 bbls/day, gas volumes ranging from 2000–10000 m3/day, and water cut ranging from 0–100%, basically covering all metering scenarios and well conditions in the field.

Combined with the existing metering and verification vehicle, tipping bucket meter, mass flow meter, vortex flowmeter and other equipment, the well station gas-liquid two-phase accuracy comparison and verification in a flexible manner can be carried out.

Among the 12 devices used for liquid phase validation, five had relative errors below 3% and seven had relative errors below 5%. During the period, a one-month continuous monitoring comparison was done in X-121, and the comparison data are shown in Fig. 2. The daily production trends measured by the three-phase flowmeter and mass flowmeter were basically the same, and the performance and stability of the equipment were verified.

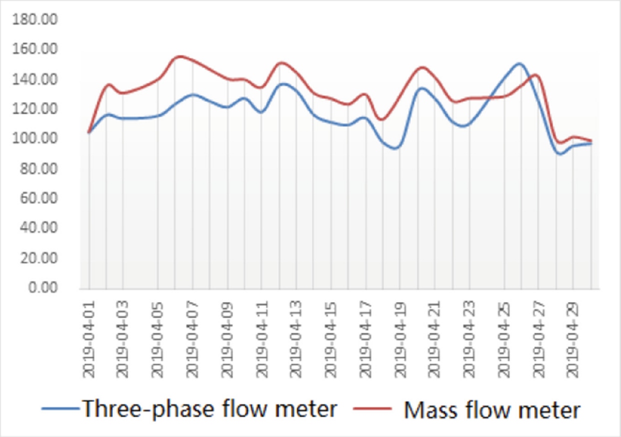


**Fig. 2.** Comparison of daily fluid production from wells at station X-121

The gas phase verification results of the 10 devices showed that the relative errors of the devices were all less than 5%. Moreover, the water content validation of 8 of the above devices showed good performance. Specifically, the absolute error of 6 devices are less than 3%, and the absolute error of 2 devices is less than 5%.

The test results show that the accuracy of the new three-phase flowmeter in the three measurement indexes of liquid volume, gas volume and water content can meet the needs of oilfield production, while combined with its higher stability can fully meet the actual production needs of the oilfield.

## 5 Conclusions

(1) The new online three-phase flowmeter has a novel design and reliable performance, and supports second-level real-time online access to oil, gas and water flow data functions, which can help oilfield users more effectively determine the trend and changes in oil and gas well production and provide a reliable basis for oilfield production management.
(2) Three phase flowmeter through the continuous field test test, find the problem, put forward solutions, and constantly improve the reliability and measurement accuracy of the product, and gradually completed the upgrade from the first generation to the second generation. Achieved a typical domestic oil field 0–100 tons / day flow detection, comprehensive coverage from high gas content low water content to low gas content high water content conditions, and the detection accuracy from 15% error rate to within 10% of the error rate.
(3) Field accuracy verification results show that the new online three-phase flowmeter can fully meet the demand of oilfield production in terms of the accuracy of liquid volume, gas volume, water content and other parameters measurement. In addition, the product has high stability and the ability to work under complex conditions, and can complete the flow measurement in various conditions in the oilfield, which is suitable for expanding applications in oil fields at home and abroad.

**Acknowledgments.** The project is supported by Science and Technology Foundation of CNPC, Construction of real-time on-line three-phase flowmeter platform - Research and development of integrated well dumping device based on real-time on-line three-phase flowmeter (Number 2017GC026-KY2022YF5018).

## References

1. Chen, B., Min, L., Zhang, Y., Liu, X.: The multi-well selector integrated based on real-time online three phase flow meter. Instrument **27**(12), 5–7 (2020)
2. Shi, P., Li, H., Ouyang, X., Liu, C.: Research on three phase flowmeter automatic swith measurement. Autom. Panorama **38**(09), 80–83 (2021)
3. Zhang, R., Liu, T., Yang, M., Dang, F.: Development analysis on the multiphase flow meter. PI, **23**(5), 30–33 (2009)
4. Fang, L., Jiang, Q., Zhang, T., Xu, Y.: Oil-gas-water three-phase flowmeter based on simple separating. Acta Metrologica Sinica **2008**(05), 445–448 (2008)
5. Mu, N.: Study on three phase measurement of low producing oil well. Northeast Petroleum University (2013)
6. Liu, S.: Research on influencing factors of measurement accuracy and structural parameter optimization of three-phase flowmeter. Harbin Institute of Technology (2013)

7. Wang, Y.: Discussion on oil-gas-water three-phase flowmeter based on simple separation method. Chem. Ind. Manage. **07**, 209 (2016)
8. Liu, Y.: Design of real-time measurement and control system for 3D sand filling physical simulation device. China University of Petroleum (East China) (2012)
9. Zhou, X.: Oil well measurement status and analysis of satellite platform in Chengdao oilfield. Neijiang Sci. Technol. **33**(02), 129–130 (2012)
10. Li, G., Ji, W., Jin, C., Sun, X., Meng, B.: Research on three-phase non-separation flow measurement technology. Oil Gas Field Surf. Eng. **05**, 20–21 (2008)
11. Sun, H., Zhao, H., Zhou, F.: The latest progress of measurement technology of oil-gas water three-phase flow. Oil Gas Storage Transp. **2002**(03), 31–37+59–6
12. Wang, Z., Han, C.: Research on XL-1 oil-gas water three-phase flow meter. Pipeline Technol. Equipment **03**, 37–39 (1996)

# Research on Intelligence Logging Interpretation Technology and System Based on Standard Big Data Platform

Ting-ting Li(✉), Hong-shu Zhang, Dao-jie Cheng, Ke Huang, Wen-mao Yu, and Hao Chen

China Petroleum Logging CO. LTD, Xi'an 710077, Shaanxi, China
569710061@qq.com, {zhanghongshucq,chengdj_cpl,huangke_cpl, zycjyuwm}@cnpc.com.cn, chenhao@cnlc.cn

**Abstract.** "Massive" logging data assets, due to their insufficient storage methods and normalization, cannot be quickly and accurately called up, become a "data island", so that their value has not been fully explored. The current application scope of artificial intelligence is focused on single method research, with few system applications. However, intelligent interpretation requires the use of a large amount of logging data and related standard data. Based on a large number of documents related to large logging database and logging artificial intelligence, starting with supervised, unsupervised and semi-supervised intelligent algorithms, this paper expounds the application status quo and applicability of intelligent logging interpretation technology through machine learning for conventional logging lithology identification, automatic layering, sedimentary microfacies identification and reservoir identification. This paper briefly introduces the application status quo of logging data governance and mining technology. This paper summarizes the process of intelligent interpretation method, as well as the intelligent logging interpretation method and system based on a physical model under a standard big data platform. This paper discusses the existing problems in intelligent logging interpretation and evaluation and the feasible development direction of future research.

**Keywords:** Artificial intelligence · Algorithm · Lithology identification · Intelligent logging interpretation

## 1 Introduction

Geophysical logging is a method to measure geophysical parameters by using geophysical characteristics such as electrochemical characteristics, electrical conductivity, acoustic characteristics and radioactivity of rock strata. During oil drilling, logging must be carried out after drilling to the designed well depth, so as to obtain various petroleum geology data and engineering technical data as the original data for completion and development of oilfields, which is called completion logging or open hole logging. All logging after casing running in oil wells or during production are generally called production logging. The development of logging has generally experienced four stages such as analog logging, digital logging, numerical control logging and imaging logging.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 144–157, 2024.
https://doi.org/10.1007/978-981-97-0272-5_12

Logging data processing and interpretation needs interpretation experts with rich regional geological experience, but even in the same area, the interpretation results of different experts are different. With the continuous development of interpretation and evaluation software, various logging interpretation software with complete functions has emerged at home and abroad, which has improved the accuracy and level of logging interpretation, promoted the progress of interpretation technology and solved some difficult problems of logging interpretation and evaluation. With the rise of big data and artificial intelligence, faced with more and more complex reservoir interpretation problems, industry experts have also realized the urgency of developing intelligent interpretation and evaluation systems and began to explore and study in this field [1].

Major international oil companies and service companies are also adjusting their development strategies, making continuous efforts in the fields of data science and artificial intelligence, and developing their own intelligent interpretation and evaluation systems. For example, Schlumberger has built its own intelligent logging processing and interpretation platform. Major oil companies and IT companies have cooperated in the field of intelligent application, and formed joint strategic research teams such as Shell + Microsoft and ExxonMobil + Microsoft, thus making many beneficial explorations in the construction of big data platform for oil and gas exploration and development, the creation of ecological environment for data sharing, and the improvement of data processing and interpretation quality, showing the great development potential of big data and artificial intelligence technology in the oil and gas industry [2].

The information abundance of data interpretation determines the application depth and breadth of logging technology, where the important link depends on the development of interpretation methods and software. As the largest producer and user of logging data in China, CNLC faces the problems such as diverse logging data types, strong professionalism and complex data operation [3].

Since 2010, China National Logging Corporation has been committed to the construction of a unified logging database with reference to the architecture of the CNPC's dream cloud platform for exploration and development. The construction of logging data resources has gone through three stages such as data management, data sharing and data application. CNLC has completed the centralized storage and management of scattered data, data sharing for individual applications and data mining applications. This has provided the required data for all kinds of professional software, helped the interconnection between systems and services, and supported the convenient sharing of data and results. On the basis of predecessors' research results, CNLC has established a data lake based on standard big data platform, followed the road of integration of logging interpretation software, realized the organic combination of three key elements: data, algorithm and scenario, improved the efficiency of logging analysis, and promoted the transformation of logging interpretation from single well interpretation to multi-well evaluation and reservoir analysis [4–8].

## 2 Data Governance Based on Logging Big Data Platform

Geophysical logging, as the "eye" of deep formations, has the advantages such as many methods, high resolution and large amount of information, and can provide continuous and accurate in-situ physical parameters such as electrical, acoustic and nuclear parameters for reservoir evaluation. After the professional reorganization of CNLC, the logging data of 16 oil and gas fields have many types and different standards, leading to extremely complex logging data; in addition, these data cannot be directly applied, forming an "island" of data, which urgently needs data governance. Focusing on three major tasks: automatic logging data sorting technology, automatic data flow technology of logging data: sorting-data governance-data warehousing and logging big data analysis technology, CNLC has carried out logging data governance work to transform unstructured data into structured data.

In order to reduce manual participation as much as possible and improve the degree of automation in the process of data governance, a multi-source heterogeneous data governance architecture based on semi-supervised learning algorithm is very suitable for logging data governance (see Fig. 2). Its basic idea is to integrate heterogeneous data describing the same entity in the real world from different data sources into structured data. The specific process includes four parts such as information extraction, pattern matching, data matching and data fusion (see Fig. 1).The actual results show that the architecture can not only effectively solve the "data island" state, but also significantly improve the data quality with as little manual participation as possible. After governance, the data call is more convenient so as to meet the requirements of different logging interpretation and geological application [14, 40–43].

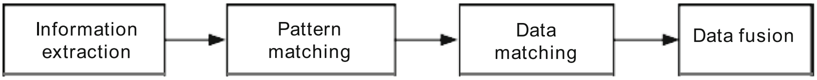


**Fig.1.** Multi source heterogeneous data governance scheme.

CNLC had formulated logging data warehousing specifications and logging data management specifications respectively for new wells and old wells. They standardize the storage file range, file format and file naming method of original logging data and result data and fully consider the current situation of the original logging database, Based on LEAD software, CNLC has developed a tool for automatically sorting, naming and exporting data. With the goal of building the brand of logging companies CNLC has unified the drawing style and formulated complete sets of drawing templates, headers and results table examples for conventional combination logging, imaging logging, production logging and engineering logging, unified curve line type, name and section filling; and standardized the symbols of logging interpretation conclusion and lithology. Moreover, CNLC has unified the mapping specifications and standardized the naming and dimensions of logging original logging curves and result curves. CNLC has established the standard name and used name, and directly managed the old well data in the original logging database to the big data platform. The new well data generated in real time is directly uploaded to the big data platform through the integrated application system.

The scope of logging data governance is defined in five logging types, which mainly include well basic information, logging curve data, map type data and table type data. It mainly evaluates the "six properties" of the data after governance, namely accuracy, completeness, standardization, uniqueness, consistency and timeliness. In view of the great differences in data storage formats of historical logging data, the non-uniform data

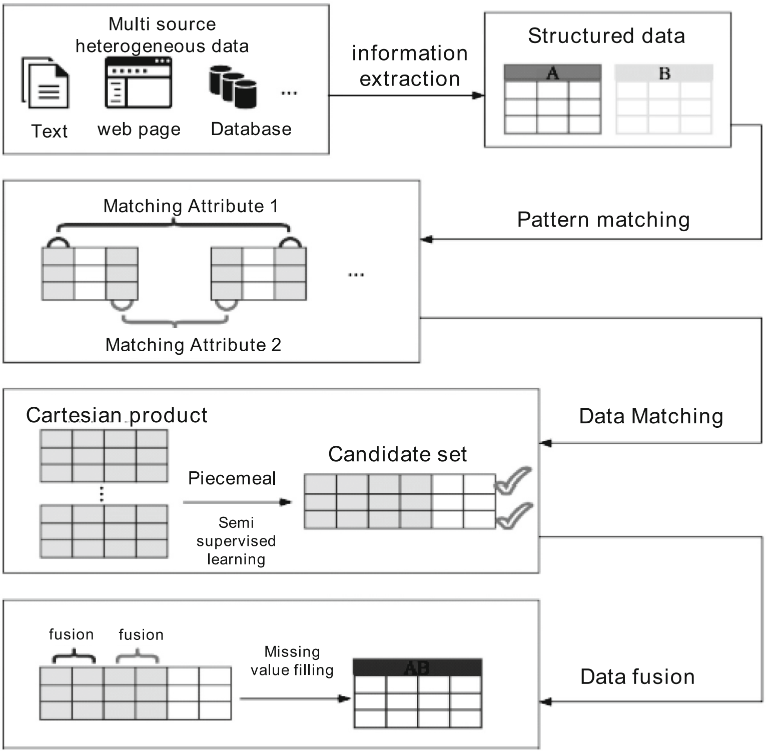


**Fig.2.** Multi-source heterogeneous data governance architecture based on semi-supervised learning (according to Wei-xiong Rao, et al.).

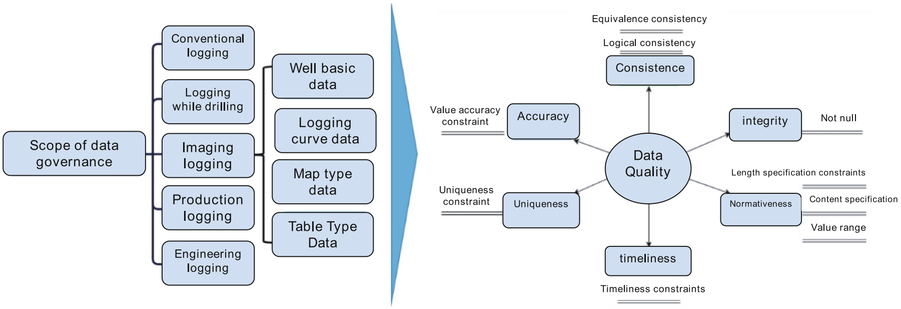


**Fig.3.** Scope of logging data governance and evaluation of "six properties" after logging data governance.

standards, and large data governance workload, there are 16 data formats and various variants of logging data achievement files, and so on, relevant decompilation tools and warehousing tools are correspondingly matched. At present, the governance of logging data of hundreds of thousands of wells has been completed, which has laid a data foundation for artificial intelligence based logging interpretation (see Fig. 3).

## 3 Present Situation and Applicability of Intelligent Logging Interpretation

The development of artificial intelligence has a long history, and it is a science based on computer technology. In logging, most machine learning models are shallow learning, such as linear classifier, BP neural network, logistic regression, K-Means clustering, support vector machine, principal component analysis, Gaussian mixture model, gradient thruster and so on. There is usually only one hidden layer in these shallow learning structures, that is, only one nonlinear feature extraction layer. Shallow learning can only be effective for some simple or limited problems in general, but it is obviously at a disadvantage in the face of complex and huge data [9].

The intelligent logging interpretation method is mainly a deep learning algorithm. Deep learning network is an extension of traditional artificial neural network. Because of its multiple hidden layers, the deep learning network can realize the mapping transformation from low-dimensional space to high-dimensional space through multi-layer nonlinear transformation, thus distinguishing complex input data features in high-dimensional space and realizing the identification and classification of complex input information. According to the characteristics of an algorithm learning task, it can be divided into supervised learning, unsupervised learning and semi-supervised learning. According to its function, supervised learning can be divided into regression and classification. Regression is to predict the occurrence probability of an object, and classification is to classify the pattern class attribution of an object. According to whether there are labels in the input data, it can be divided into supervised learning and unsupervised learning. Unsupervised algorithms, such as clustering algorithm and dimension reduction algorithm, such as fisser discriminant method, are effective in identifying complex oil-water layers. Logical regression, support vector machine, proximity regression and decision tree algorithms in supervised algorithms have high accuracy in identifying complex lithology. On the basis of computer algorithm, semi-supervised algorithm is added with human experience, so that the clustering effect is significantly improved by using a small amount of labeled data and a large amount of unlabeled data. The semi-supervised classification task combined with unsupervised learning dimension reduction method can improve the classification effect of supervised learning under the condition of insufficient labels. Swarm intelligent optimization algorithm is an algorithm that combines the behaviors of animals such as foraging and avoiding obstacles, including bat algorithm, ant colony algorithm and firefly algorithm, and is used to study the wave impedance inversion method [46–53].

Industry scholars have used vector machine, neural network, fuzzy recognition and traditional decision tree methods to identify lithology, and achieved good application results [10–12].Wang Hua et al. deeply analyzed the applicability of applying artificial intelligence in logging data processing and interpretation from the traditional data modeling method and machine learning algorithm in geophysical logging field. Chen Xi et al. expounded that artificial intelligence based logging interpretation is feasible from three cores such as data model, physical simulation algorithm and artificial intelligence based logging ecology, which can help logging analysts solve deeper geological problems [1, 13].

# 4 Artificial Intelligence Based Interpretation of Logging Data

## 4.1 Lithology Identification

The existing lithologic identification method is mainly to calibrate the logging curve through a small number of logging cores, and use the obtained logging curve data to identify the lithology of the whole interval. Intelligent interpretation method is combined with logging processing and interpretation to identify lithology. Its general idea is to select logging curves sensitive to lithology identification as input curves based on core data and lithology sensitivity analysis of logging curves, so as to realize lithology identification based on intelligent algorithm.

Decision Tree Algorithm is a kind of supervised learning. According to the weight of logging parameters in clastic rock research area in lithology identification model, the sensitivity of each parameter to lithology change is determined, so as to identify lithology. The decision tree method of C5.0 has effectively improved the accuracy of lithology identification. Decision tree algorithm also has high accuracy in identifying complex carbonate rocks. For the model with huge data, XGBoost algorithm can be applied, which adopts multithreading and distributed computing methods, greatly shortening the training time. It has a good recognition effect on limestones and dolomites, followed by argillaceous limestones, dolomites and argillaceous dolomites, and the recognition rate of calcareous dolomites is low. Boosting Tree algorithm can also effectively determine the lithology of complex glutenites [15–18].

The random Forest Algorithm, which shows great advantages in thin layer identification, has strong generalization ability, insensitivity to feature loss, fast training speed and simple implementation. Based on the lithology sensitivity analysis of logging curves, a suitable logging curve is selected as the input curve, and the lithology identification model of complex carbonate rocks is established by using random forest algorithm, which is accurate for lithology identification [10].

Another advantage of random forest algorithm is lithology identification of volcanic rocks. The lithology of volcanic oil and gas reservoirs is changeable, so it is difficult to identify them accurately by conventional methods. Different types of volcanic rocks include volcanic breccia and fused breccia. Lava mainly includes basalt, andesite, dacite and rhyolite. Due to their differences in chemical composition, mineral composition and physical properties, there are some changes in their corresponding logging response characteristics, thus identifying lithology [21].

Principal Component analysis (PCA) is one of unsupervised learning, and the key to identification is to convert the comprehensive response characteristics of various logging curves to the principal component of prominent lithology, so the identification accuracy of alternate thin layers of volcanic rocks or shale with complex lithology is high [20, 22]. By combining BP neural network lithologic prediction model and Dropout mechanism, Dropout-BP neural network combines conventional logging parameters, upgrading the conventional two-parameter crossplot to a multi-parameter neural network, and integrating the composition, structure and electrical properties of volcanic rocks to carry out lithologic prediction, which is more effective [23].

Data mining method of logging lithology identification based on emergent self-organizing mapping. Large-scale neurons and borderless torus mapping are used, visualized by U matrix, and finally clustered and classified by manual interaction. This method can effectively find hidden patterns in high-dimensional data, and is especially suitable for logging identification of complex lithology [24].

The application of multivariate statistical algorithm needs to preprocess logging data, including logging parameter selection, logging data normalization and dimensionality reduction. Its application effect is good [25], but it can only solve the simple linear relationship problem.

By constructing the technology of recovering the missing core picture information and combining with migration learning, mine lithology is identified. The corresponding core sample information is automatically synthesized from the logging curve data of non-coring wells, and the existing data is automatically learned and analyzed by using migration learning technology, so that the core sample information of coring wells is migrated to non-coring wells, and a logging lithology identification model aided by core samples is established. The establishment of intelligent identification model of cores based on migration learning is helpful to improve the accuracy of lithology identification of oil and gas reservoirs with complex cores, and logging curves are used to predict lithology quickly and accurately [26].

In order to integrate the algorithms and modules into the unified software, the data resource Lead software has been developed by CNLC, which includes reservoir parameter calculation module based on conventional logging data, single porosity calculation module based on clastic rock, CRA module based on carbonate rock calculation and CLASS module based on lithology classification. Its advantage is that it can choose the appropriate calculation model according to the background of different regions, which is convenient for rapid processing and interpretation of logging data. But the model coverage is not comprehensive enough.

### 4.2 Automatic Layering and Identification of Reservoirs

There are three main guiding ideas for automatic layering: (1) variance analysis of logging values and finding inflection points and half amplitudes on curves. The guiding ideology of variance analysis is that the intra-layer difference is small and the inter-layer difference is large. At the same time, the inflection point and half amplitude point are found on the logging curve by differential and slope extreme point. (2) According to the logging data, judge the rock attributes or calculate the membership degree of rocks, and merge the same lithology, so as to realize layering (see Fig. 4). (3) Divide strata by flow unit method based on fluid properties. In actual interpretation work, automatic layering is carried out according to the priority order of fluid > lithology > curve [29, 30].

Three kinds of methods: mathematical statistical methods include intra-layer difference method, ordered cluster analysis, extreme variance clustering method and change point analysis method (least square method and maximum likelihood estimation method); Non-mathematical statistical methods include activity function method and wavelet transform method; Artificial intelligence methods include cluster analysis, fuzzy mathematics and neural network methods. These methods have their respective advantages and disadvantages.

Mathematical statistical method is strict in mathematics, which can keep the uniformity inside the rock strata, the difference between the rock strata is great, and the calculation amount is large. In addition, It has a very high requirement for the one-to-one correspondence between logging information and geological information. If it can't be achieved (in fact, it can't be completely achieved), the layering result is a perfect mathematical result, which is not easy to meet the requirements of geological application. Among the non-mathematical statistical methods, the activity function method has good application effect and can quickly identify various types of curves; Wavelet transform can simulate the artificial interpretation process of "from coarse to fine, layering step by step" through multi-scale analysis, so as to avoid layering on a visual level and being unable to distinguish between the local and overall information of strata [27].

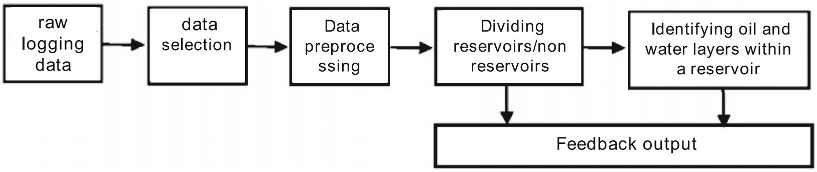

raw logging data
data selection
Data preproce ssing
Dividing reservoirs/non reservoirs
Identifying oil and water layers within a reservoir
Feedback output

**Fig.4.** Technical route of automatic layering and identification with logging curves.

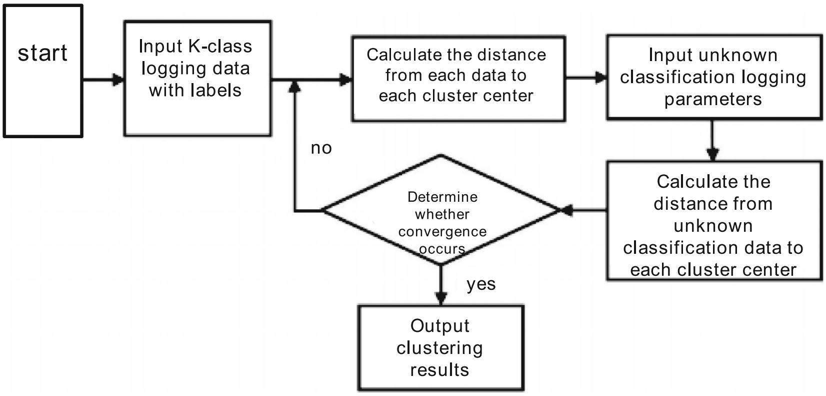

start
Input K-class logging data with labels
Calculate the distance from each data to each cluster center
Input unknown classification logging parameters
no
Determine whether convergence occurs
Calculate the distance from unknown classification data to each cluster center
yes
Output clustering results

**Fig.5.** Flow chart of multi-granularity clustering method (according to Ji Qingqing).

In the application of artificial intelligence methods, the multi-granularity clustering algorithm with good effect belongs to supervised learning (see Fig. 5). This intelligent algorithm can quickly and accurately solve various classification problems, extract the characteristics of different layered logging curves by learning standard logging curves and layering results, and then identify oil-water layers on the basis of dividing reservoirs. In the case of standard big data, firstly, the original logging curves are analyzed by principal component analysis, and then the relationship between each original logging curve and principal component is analyzed by principal component load matrix, and then the logging curves used for automatic layering with logging curves are selected [28, 44, 45].

The other is a knowledge-driven neural network reservoir evaluation model (KPNFE) based on the knowledge map of reservoir logging. Its functions mainly include: (1) multi-dimensional and multi-scale extraction of characteristic parameters that describe oil and gas reservoirs in detail; (2) The entities, relationships and attributes associated with these characteristic parameters are represented as vector characteristic graphs by graph embedding technology; (3) Realizing intelligent identification of oil and gas reservoirs; (4) Organically integrate expert knowledge into intelligent computing, and establish an evaluation system and optimization algorithm for potential layer recommendation [32]. The KPNFE model inherits and promotes expert knowledge and experience, effectively solves the problem of robustness in oil and gas reservoir identification, and its calculation results are highly interpretable and accurate, and it is an effective method for re-logging evaluation of old wells in old areas with high efficiency and high quality.

### 4.3 Sedimentary Microfacies Identification

The traditional method of identifying sedimentary microfacies is achieved manually by geologists according to their own knowledge and experience. This manual interpretation is subjective and time-consuming, and may introduce human bias. The method of identifying sedimentary microfacies based on logging curves usually includes three steps: logging curve layering, feature extraction and classification [31].Typical classification algorithms include Bayesian criterion, linear discriminant analysis, fuzzy logic, convolutional neural network method, K nearest neighbor algorithm, SVM, ANN and so on. The process of depth learning method based on logging curve is as follows: (1) data preprocessing: (2) data marking and division: (3) model training: (4) model verification. However, due to its own limitations, a single intelligent method is difficult to complete the task of sedimentary microfacies identification alone.

Convolutional neural network method takes into account the morphological change characteristics of logging curves in depth direction and the need to integrate the three steps of curve layering, feature extraction and classification. Aiming at the multi-scale and time series of logging curves, a logging sedimentary microfacies identification model, Improved U-net, with multi-scale characteristics constraints has been established, which can well identify distributary channel, channel side margin and distributary bay with different scales. KD-SegCaps, a logging sedimentary microfacies identification model with time series constraints, can well identify sedimentary microfacies such as sand flat, sand mud flat and mud flat [31]. Using DMC-BiLSTM, an intelligent identification method of sedimentary microfacies based on feature construction (DMC) and bidirectional long-term and short-term memory network (Bilstm), the geological trend characteristics, median filtering characteristics and clustering characteristics have been constructed. Compared with the convolutional neural network method, this method is helpful to extract the hidden features of logging curve sequence, and has better recognition performance for sedimentary microfacies such as distributary bay, front sheet sand, distributary channel, estuary bar and channel side maigin [39].

## 5 Intelligent Logging Interpretation Method Process and Data Architecture

Intelligent logging interpretation integrates all kinds of deep learning algorithms combining the characteristics of logging interpretation business, so that intelligent algorithms are integrated with traditional logging interpretation concepts. Its steps are intelligent model training, model combination and automatic recommendation (see Fig. 6).

The Digital Reservoir Research System (RDMS) pioneered by Changqing Oilfield is divided into four layers such as data layer, data link, support layer and application layer. Functionally, it includes five platforms: basic management, data service, collaborative research, decision support and cloud software [1, 13, 34, 35]. Inspired by this model, CNLC has established a big data ecology based on logging data lake (see Fig. 7). Taking logging data as the main body, CNLC has built up a regional lightweight lake by gathering logging data at home and abroad. Data are transmitted to various professional libraries by means of automatic collection of the Internet of Things and manual standardized collection, and then merged into the data lake after cleaning and processing. Real-time data and video data of industrial control are stored nearby [4]. CNLC has studied key technologies such as data integration and professional software interface, and developed and integrated exploration and development business model, multi-source data of oil and gas reservoirs, multidisciplinary professional software and online analysis tools. Realizing the coupling and integration of professional software, intelligent application and data lake. In the application scenario, data loading can be completed. On the basis of core analysis, the interpretation conclusion has been re-recognized, the logging characteristic values of the target horizon of each well have been marked, and the sample data have been submitted in batches by layers. In addition, these data are stored in the local computer work area, and can be adjusted and updated to the sample library at any time. The mode from big data platform to data lake plus interface can meet the requirements of different logging geological structure analysis scenarios.

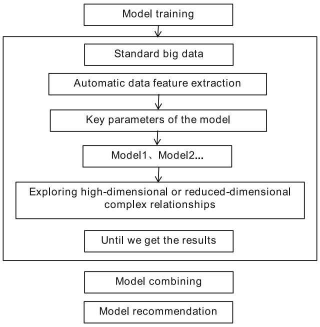

Model training
Standard big data
Automatic data feature extraction
Key parameters of the model
Model1、Model2...
Exploring high-dimensional or reduced-dimensional complex relationships
Until we get the results
Model combining
Model recommendation

**Fig.6.** Intelligent Interpretation Method Process.

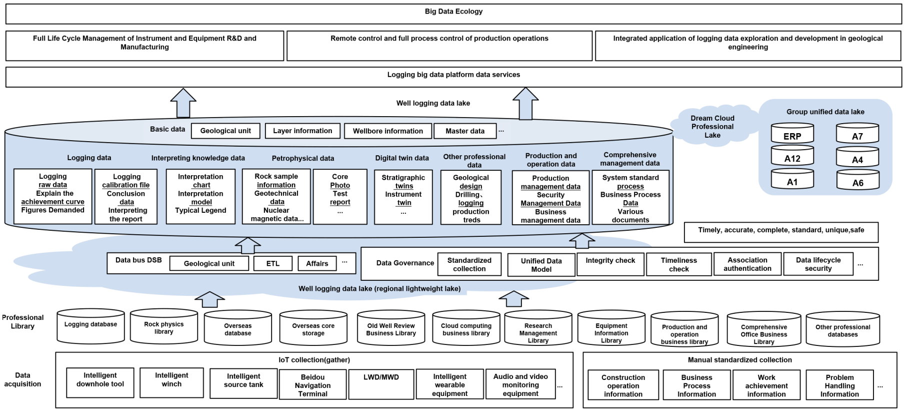

Big Data Ecology
Full Life Cycle Management of Instrument and Equipment R&D and Manufacturing
Remote control and full process control of production operations
Integrated application of logging data exploration and development in geological engineering
Logging big data platform data services
Well logging data lake
Basic data
Geological unit
Layer information
Wellbore information
Master data
...
Logging data
Interpreting knowledge data
Petrophysical data
Digital twin data
Other professional data
Production and operation data
Comprehensive management data
Logging raw data Explain the achievement curve Figures Demanded
Logging calibration file Conclusion data Interpreting the report
Interpretation chart Interpretation model Typical Legend
Rock sample information Geotechnical data Nuclear magnetic data...
Core Photo Test report ...
Stratigraphic twins Instrument twin ...
Geological design Drilling、logging production treds
Production management data Security Management Data Business management data
System standard process Business Process Data Various documents
Dream Cloud Professional Lake
Group unified data lake
ERP
A7
A12
A4
A1
A6
Timely, accurate, complete, standard, unique,safe
Data bus DSB
Geological unit
ETL
Affairs
...
Data Governance
Standardized collection
Unified Data Model
Integrity check
Timeliness check
Association authentication
Data lifecycle security
...
Well logging data lake (regional lightweight lake)
Professional Library
Logging database
Rock physics library
Overseas database
Overseas core storage
Old Well Review Business Library
Cloud computing business library
Research Management Library
Equipment Information Library
Production and operation business library
Comprehensive Office Business Library
Other professional databases
Data acquisition
IoT collection(gather)
Intelligent downhole tool
Intelligent winch
Intelligent source tank
Beidou Navigation Terminal
LWD/MWD
Intelligent wearable equipment
Audio and video monitoring equipment
...
Manual standardized collection
Construction operation information
Business Process Information
Work achievement information
Problem Handling Information
...

**Fig.7.** Intelligent Logging Interpretation Data architecture diagram Based on Standard Big data Platform(CNLC).

## 6 Summary and Prospect

There are many kinds of intelligent interpretation methods, and different methods have their own advantages and disadvantages in lithology identification, automatic stratification, hydrocarbon reservoir identification and sedimentary microfacies identification of clastic rocks, complex carbonate rocks, shale and volcanic rocks. Through model training, the optimal method can be obtained so as to improve the interpretation accuracy and efficiency of complex reservoirs.

Big data is the foundation of intelligent interpretation. In practice, a high variable dimension may not have high analytical accuracy, and sometimes it may even have the opposite effect. Through the logging data management and data quality evaluation of the big data platform, the logging data of hundreds of thousands of wells have been managed to ensure the accuracy, completeness and standardization of the data and facilitate data call. In the application scenario, the interpretation conclusion has been re-recognized on the basis of core analysis. The mode from big data platform to data lake plus interface can meet the requirements of different logging geological structure analysis scenarios.

Generally speaking, the infrastructure layer realizes IOT perception and resource support, the data sharing layer realizes data entering the lake and comprehensive management, the middle platform layer builds shared and reused data and business service capabilities, and the application layer builds lightweight and agile intelligent application scenarios. It has realized the transformation of production and operation from man-machine combination to intelligent cooperation, business management from process-driven to data-driven, and business decision-making from experience management to intelligent analysis, thus building a digital logging ecology and building a digital enterprise.

## References

1. Hua, W., Yushun, Z.: Research status and prospect of artificial intelligence in logging data processing and interpretation [J]. Well Logg. Technol. **45**(4), 345–356 (2021)

2. Yong, Y.: Application progress of big data & AI technologies in exploration and development of Shengli Oilfield [J]. Petrol. Geol. Recov. Effi. **29**(1), 1–10 (2022)
3. Ailin, J., Jianlin, G.: Key technologies and understandings on the construction of smart fields. Pet. Explor. Dev. **39**(1), 118–122 (2012)
4. Yujiang, S., et al.: Development and application of intelligent logging interpretation system based on big data [J]. China Petrol. Explor. **26**(2), 113–126 (2021)
5. Yong, Y., et al.: Research on construction of data interlinked lakes of E & P dream cloud [J]. China Petrol. Explor. **25**(5), 82–88 (2020)
6. Zhiyong, L., et al.: Technology research and construction scheme of unified data lake [J]. Telecommun. Sci. **37**(1), 121–128 (2021)
7. Zhou, J., et al.: Research and application of well logging database system [C]. In: China Petroleum and Petrochemical Enterprise Information Technology Exchange Conference, pp. 393–398 (2016)
8. Zhou, J., et al.: Construction and application of unified logging database [J]. Well Logg. Technol. **46**(6), 757–761 (2022)
9. Fan, H.: Research on intelligent recognition of indicator diagrams based on artificial intelligence algorithms [D]. In: Henan University of Science and Technology, Luoyang (2019)
10. Wang, Q., et al.: Identification of complex carbonate lithology based on random forest algorithm [J]. Chin. J. Eng. Geophys. **17**(5), 550–558 (2020)
11. Teng, M.: Research on Wave Impedance Inversion Method Based on Improved Firefly Algorithm [D]. Jilin University, Jilin (2021)
12. Sun, Y., et al.: Logging identification of complex carbonate rock lithology based on XGBoost algorithm [J]. Lithological Reservoirs, **32**(4), 98–106 (2020)
13. Cheng, X., Song, X., Li, G.: Design and implementation of ecological clouds for artificial intelligence logging driven by big data and algorithms [J]. Well Logg. Technol. **45**(3), 233–239 (2021)
14. Rao, W., et al.: Multi-source heterogeneous data governance based on semi-supervised learning [J]. J. Tongji Univ. (Natural Science), **50**(10), 1393–1404 (2021)
15. Miao, T., et al.: Lithology identification method for clastic rock based on data mining technology and its application [J]. Complex Reservoirs, **14**(1), 39–44 (2021)
16. Rui, W., Xiaomin, Z., Lichang, W.: Identifying carbonate rock lithology by data mining [J]. Well Logg. Technol. **36**(2), 197–201 (2012)
17. Kai, J., et al.: A logging lithology identification model based on boosting tree algorithm [J]. Well Logg. Technol. **42**(4), 395–400 (2018)
18. Wang, Z., Liu, J., Ren, L.: The method for lithology classification in geophysical well logging based on the k-means dynamic clustering analysis [J]. J. East China Inst. Technol. (Natural Science Edition), **32**(2), 152–156 (2009)
19. Mao, R., et al.: Identification of mixed rock lithology based on lithology scanning logging: taking the Fengcheng formation of Mahu sag as an example [J]. Xinjiang Petrol. Geol., **43**(6), 743–749 (2022)
20. Liu, Y., et al.: Application of principal component analysis method in lithology identification for shale formation [J]. Fault-Block Oil Gas Field, **24**(3), 233–239 (2017)
21. Kai, L., et al.: Intelligent identification and prediction of volcanic rock lithology based on machine learning [J]. Spec. Res. **29**(1), 38–45 (2022)
22. Li, S., et al.: Lithology identification of carboniferous volcanic rocks in Xiquan area, eastern Junggar Basin [J]. Lithological Reservoirs, **33**(1), 258–266 (2021)
23. Yiming, H., et al.: Identification of volcanic rock lithology using neural networks based on conventional logging data: a case study of the changling fault depression in the southern songliao basin [J]. World Geology **40**(2), 408–418 (2021)

24. Haifeng, G., et al.: Data mining based on emergent Self-organizing mapping: a new method for logging lithology identification [J]. J. Petrol. Nat. Gas (J. Jianghan Petrol. Inst.) **31**(2), 67–70 (2009)
25. Li, Z., et al.: Application of data mining methods in logging lithology identification [J]. Fault Block Oil Gas Field, **26**(6), 713–718 (2019)
26. Ren, Y., et al.: Intelligent identification technology for lithology based on comprehensive consideration of core images and logging curves [J]. Tech. Appl. 78–80 (2021)
27. Bo, X., et al.: Review and prospect of automatic layering methods with logging curves [J]. Prog. Geophys. **25**(5), 1802–1810 (2010)
28. Ji, Q., et al.: Automatic layering recognition method for logging curves based on multi-granularity clustering [J]. Chin. High Technol. Lett. **30**(12), 1215–1224 (2020)
29. Xianmin, Z., Jianhua, W.: Artificial intelligence of the lithologic auto-recognition from digital well-logging data in engineering [J]. J. Tianjin Univ. **34**(5), 633–635 (2001)
30. Hui, Y., et al.: Application of wavelet transform characteristics of logging curves in automatic stratification. J. Geophys. **43**(4), 568–573 (2000)
31. Yongxiang, H.: Intelligent Identification Method of Sedimentary Microfacies from Logging Curves [D]. University of Electronic Science and Technology of China, Chengdu (2021)
32. Guoqiang, L., et al.: Construction of well logging knowledge graph and intelligent identification method of hydrocarbon-bearing formation [J]. Pet. Explor. Dev. **49**(3), 502–512 (2022)
33. Li, H., et al.: Research on data mining methods for logging evaluation of complex reservoirs [J]. J Acta Petrolei Sinica, **30**(4), 542–548 (2009)
34. Yang, H., et al.: Construction and application of an integrated information platform for reservoir research and decision-making [J]. China Petrol. Explor. **20**(5), 2–8 (2015)
35. Tan, M., et al.: Progress in research on committee machine logging interpretation driven by multiple source data by the commission [J]. Geophys. Prospect. Petrol. **61**(2), 233–239 (2022)
36. Di, Q., et al.: Intelligent steering drilling technology based on cloud big data [J]. J. Eng. Geol. **29**(1), 162–170 (2021)
37. Changchun, Z., et al.: Research on ROP prediction model based on fusion of Feature selection algorithm [J]. Drill. Eng. **49**(4), 233–239 (2022)
38. Jianguo, C., Wei, M., Ye, L., et al.: Research of sandstone reservoir physical properties estimation based on Elman neural networks with hybrid dimensionality reduction [J]. Sci. Technol. Eng. **14**(3), 24–28 (2014)
39. Renze, L., et al.: Intelligent identification of sedimentary microfacies based on DMC-BiLSTM. Geophys. Prospect. Petrol. **61**(2), 253–261 (2022)
40. Wang, X., et al.: Koko: a system for scalable semantic querying of text[C]. In: Proceedings of the VLDB Endowment, pp. 2018–2021. VLDB Endowment, Rio de Janeiro (2018)
41. Mudgal, S., et al.: Deep learning for entity matching: a design space exploration [C]. In: Proceedings of the 2018 International Conference on Management of Data, pp. 19–34. ACM, New York (2018)
42. Trivedi, R., et al.: LinkNBed: multigraph representation learning with entity linkage [C]. In: Proceedings of the 56th Annual Meeting of the Association for Computational, pp. 252–262. ACL, Melbourne (2018)
43. Konda, P., et al.: Technical perspective: toward building entity matching management systems [J]. ACM SIGMOD Rec. **47**(1), 33 (2018)
44. Amanipour, V., Ghaemmaghami, S.: Median filtering forensics in compressed video [J]. IEEE Signal Process. Lett. **26**(2), 287–291 (2019)
45. Green, O.: Efficient scalable median filtering using histogram-based operations [J]. IEEE Trans. Image Process. **27**(5), 2217–2228 (2018)
46. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning [M]. MIT Press, Cambridge (2016)

47. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neuralnetworks [C]. In: The 25th International Conference on Neural Information Processing Systems, New York, USA (2012)
48. Vincent, P., et al.: Stacked denoising autoencoders:learning useful representations in a deep network with a local denoising criterion [J]. J. Mach. Learn. Res. **11**(12), 3371–3408 (2010)
49. Hinton, G.E., Osindero, S., The, Y.W.: A fastlearning algorithm for deep belief nets [J]. Neural Comput. **18**(7), 1527–1554 (2006)
50. Wagstaff, K., et al.: Constrained K-means clustering with background knowledge[C]. In: Eighteenth International Conference on Machine Learning, San Francisco, USA (2001)
51. Basu, S., Banerjee, A., Mooney, R.J.: Active semi-supervision for pairwise constrained clustering [C]. In: Proceedings of the SIAM International Conference on Data Mining, Lake Buena Vista, USA (2004)
52. Ren, Y., et al.: Semi-supervised deepembedded clustering [J]. Neurocomputing **325**(1), 121–130 (2019)
53. Bergen, K.J., et al.: Machine learning for data-driven discovery in solidearth geoscience [J]. Science **363**(6433), eaau0323 (2019)

# Hybrid Model Based on Attention Mechanism for Production Prediction of Sucker Rod Well

Xin-yan Wang[1], Kai Zhang[1,2](✉), Li-ming Zhang[1], Cheng Cheng[1], Pi-yang Liu[2], and Xia Yan[1]

[1] College of Petroleum Engineering, China University of Petroleum (East China), Qingdao, China
zhangkai@upc.edu.cn
[2] School of Civil Engineering, Qingdao University of Technology, Qingdao, China

**Abstract.** In oilfield production, the liquid production is an important indicator for measuring the production capacity of sucker rod wells and determining reasonable production parameters. Therefore, accurate metering of liquid production in sucker rod wells holds significant importance for oilfield automation production management. This paper proposed a physical-data hybrid-driven liquid production prediction method based on the attention mechanism to improve the accuracy of sucker rod well production metering. First, a physical-driven model for measuring liquid production based on the sucker rod well dynamometer cards is established, which ensures the rationality and interpretability of predicting liquid production. Then, a ResNet-based data-driven model is established to uncover the hidden features in downhole pump dynamometer cards and oil well production data. Finally, an attention mechanism is employed to couple the physical-driven and data-driven models, facilitating the identification of crucial features for liquid production prediction. The proposed method was tested on actual production data, and the average accuracy rate reached 95.67%, which was at least 2.43% higher than other best benchmark models for production prediction, and demonstrating good prediction accuracy and stability in special operating conditions. This approach successfully fuses the physical analytical model and data mining model of sucker rod wells, ultimately enhancing the interpretability and reliability of the model, thereby promoting efficient production management in oilfields.

**Keywords:** Sucker Rod well · Production Prediction · Hybrid Modeling · Dynamometer Card · Attention Mechanism

## 1 Introduction

With the continuous growth of global energy demand and the depletion of oilfield resources, there is an urgent need to achieve sustained and stable production in oilfields through automation and intelligent production optimization management. The sucker rod well is the most commonly employed artificial lift system in the oil and gas industry, and liquid production serves as a vital metric for evaluating well productivity and determining optimal production parameters. However, automating the measurement of

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 158–173, 2024.
https://doi.org/10.1007/978-981-97-0272-5_13

liquid production in sucker rod wells is challenging due to the complexity of the process flow and high maintenance costs of traditional mechanical liquid measurement mode [1], which hinder oilfield automation management.

With the development of electronic measurement technology and the industrial Internet of Things (IoT) [2], the virtual flow measurement technology based on dynamometer cards in sucker rod wells has gradually gained wide acceptance in the oilfield industry since the 1980s [3], due to its low cost, acceptable error range, and remote operability. This physical-driven model utilizes a vibration mathematical model of the pump rod to solve for the dynamometer card, which indicates the pumping performance of the downhole pump. By analyzing the effective stroke of the pump plunger and quantifying various parameters, the effective liquid production at the wellhead can be calculated. In practical applications, the accuracy and stability of this method can be affected by the simplified system theory models and the complex working conditions of the downhole pump. In light of existing problems, various advanced approaches have been proposed. In 2013, Lyu et al. [4], proposed an interactive method for obtaining pump valve points based on prior knowledge of dynamometer cards and manual experience, which reduces the effective stroke error. In 2020, Yin et al. [5] proposed an analytical solution easily applied for predicting the behaviors of multi-tapered sucker-rod pumping systems, which provides a more precise description of the motion characteristics of the downhole pump. In 2020, Lv et al. [6] proposed a production measurement method based on quantitative analysis of fault dynamometer cards, which effectively improved the accuracy of liquid production prediction under valve leakage conditions. Nevertheless, there is an immeasurable gap between physical models based on prior information and the real world. This difference leads to inaccuracies in the liquid measurement, and further optimization is necessary to address these issues.

In recent years, artificial intelligence (AI) technology has emerged as the engine driving the "Fourth Industrial Revolution," and it has played a significant role in the digital transformation and intelligent development of the oil and gas industry. Machine learning methods, with their intelligence, simplicity, and efficiency, are widely utilized to address traditional engineering problems [7]. In 2019, Ruiz et al. [8] employed fuzzy logic (FL) and artificial neural networks (ANN) to interpolate oil well data and select the most effective features for predicting production. In 2021, Pan et al. [9] combined convolutional neural networks (CNN) and long short-term memory neural networks (LSTM) alongside attention mechanisms to forecast production with time series data derived from the oil well.

The dynamometer cards, as the most effective indicator data for characterizing the motion characteristics of sucker rod well system, has significantly improved its fault identification and liquid production measurement accuracy due to the application of machine learning models. In 2020, Peng et al. [10] employed a deep autoencoder to extract high-dimensional features from the dynamometer cards, aiming to overcome the limitations of traditional manual feature extraction methods. In 2022, Zhang et al. [11] aimed to the disadvantage that the traditional dynamometer card diagnosis needs a large number of samples, a small sample diagnosis framework based on meta transfer learning is proposed. However, whether it is machine learning or deep learning, the characteristic of these data-driven models is to explore and utilize the underlying patterns in the data. The drawback is that they often lack higher-order explanations in terms of real-world physical significance and may suffer from overfitting and limited generalization capabilities. Furthermore, it is important to note that the current methods for liquid production prediction generally lack the strong theoretical foundation provided by the measurement based on dynamometer cards.

In this work, we presented a hybrid-driven prediction model for liquid production of sucker rod wells that integrated physical and data-driven models using an attention mechanism. A mathematical model was employed to solve the dynamometer cards of the downhole pump, and then quantitative analysis was conducted on the cards to extract physical features that characterized the pump's operational state and theoretical displacement. This ensured that the hybrid model possessed reliable global characteristics. To address limitations in the physical model and quantitative analysis, while using Resnet to extract the high-dimensional features of the surface dynamometer cards. The attention mechanism is used for concentrating on effective features and reduce the impact of low-contributing and ineffective features, which guarantees the high accuracy and robustness of hybrid model.

The remaining work of the paper is arranged as follows. Section 2 introduces theoretical methods in oil production engineering and machine learning. Section 2.1 discusses the production measurement based on dynamometer cards, while Sects. 2.2 and 2.3 present the fundamental theories of the deep learning network ResNet and attention mechanism. Section 3 introduces the hybrid model for production prediction. Section 3.1 presents the detailed structure of the hybrid model. Section 3.2 elaborates on the establishment of the physical model and the steps for extracting physical features. Section 3.3 describes the modeling approach of the hybrid-driven model based on ResNet and the attention mechanism. Section 4 validates the performance of the model through comparative experiments and ablation study. Section 5 summarizes the main contributions of this paper and provides an outlook for future work.

## 2 Methodology

### 2.1 Production Measurement Based on Dynamometer Cards

**Calculation of Downhole Pump Dynamometer Card**

The surface pumping unit is connected to the downhole pump via sucker rods, enabling reciprocating motion. The displacement and load of the surface pumping unit's hanging point are recorded using a dynamometer card. However, the downhole pump is subject to various disturbances, forces, and torques, resulting in vibration or impact phenomena. Therefore, the surface dynamometer card cannot accurately depict the downhole pump's motion characteristics and operational state. Consequently, it is necessary to establish a model of the sucker rod well motion system to mathematically convert the surface dynamometer card into a downhole pump dynamometer card.

The sucker rod well motion system model is a mathematical model that describes the dynamic characteristics of a pumping unit well system. The Gibbs [12] model utilizes a wave equation with viscous damping as the fundamental differential equation to describe the dynamic behavior of the sucker rod:

$$\frac{\partial U(x,t)}{\partial t^2} = a^2\frac{\partial^2(x,t)}{\partial x^2} - c\frac{\partial U(x,t)}{\partial t} \tag{1}$$

where, $U(x, t)$ is the displacement of any cross-section $(x)$ of the sucker rod column at any given time $(t)$, **m**; $a$ is the stress wave propagation velocity, **m/s**; $c$ is equivalent damping factor, **1/s**.

The dynamic load function of the hanging point expressed by the truncated Fourier series and the displacement function of the light rod are used as the boundary conditions, and the motion equation of the cross-section of the sucker rod at any depth can be obtained by the separation variable method:

$$U(x,t) = \frac{\sigma}{2EA_r}x + \frac{\nu_0}{2} + \sum_{n=1}^{N}[O_n(x)\cos n\omega t + P_n(x)\sin n\omega t] \tag{2}$$

where, $E$ is the rod pump Young's modulus, **Pa**; $A_r$ is the rod string cross-sectional area, $\mathbf{m^2}$; $\sigma$, $\nu_0$, $O_n(x)$ and $P_n(x)$ are all Fourier coefficients.

According to Hooke's law, the time-varying dynamic load on that section can be determined:

$$F(x,t) = EA_r[\frac{\sigma_0}{2EA_r} + \sum_{n=1}^{N}[\frac{\partial O_n(x)}{\partial x}\cos n\omega t + \frac{\partial P_n(x)}{\partial x}\sin n\omega t] \tag{3}$$

where, $F(x, t)$ is the dynamic load on any cross-section at a given depth $(x)$ of the sucker rod, **N**. At time $t$, the total load on the cross-section at depth $(x)$ is equal to the sum of the dynamic load $F(x, t)$ and the weight of the sucker rods below the x-section.

A conversion example is shown in Fig. 1. The downhole pump dynamometer card exhibits a smoother and more stable shape by eliminating the deformation of the sucker rod column, rod friction, vibrations, and inertia. This will facilitate quantitative analysis of the pump dynamometer card to determine the effective stroke of the plunger $S_p$.

**Calculation of Sucker Rod Well Production**

The effective plunger stroke $S_p$ is primarily determined based on the position of the valve opening and closing points on the pump dynamometer card. Typically, the smaller displacement difference between the traveling valve switching point and the standing valve switching point is used as $S_p$. For example, in Fig. 2(a)(b)(c), the length of segment AD is considered the effective stroke, while during plunger unloading, the $S_p$ corresponds to the smaller length of segment BC.

Therefore, without considering the conditions of tubing leakage and pump leakage, the actual daily production at the wellhead of a pumping unit well can be calculated using the following equation:

$$Q = 1440\frac{\pi D_p^2}{4}S_pNB_l \tag{4}$$

where, $Q$ is the daily production rate, $\mathbf{m^3/d}$; $D_p$ is the diameter of the pump, **m**; N is the stroke number, $\mathbf{min^{-1}}$; $B_l$ is the volume coefficient of the crude oil with dissolved gas.

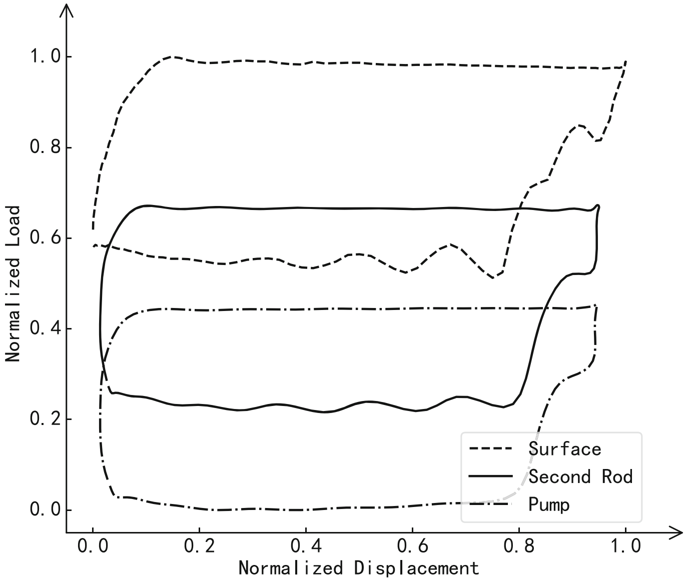


**Fig. 1.** Downhole pump dynamometer card conversion

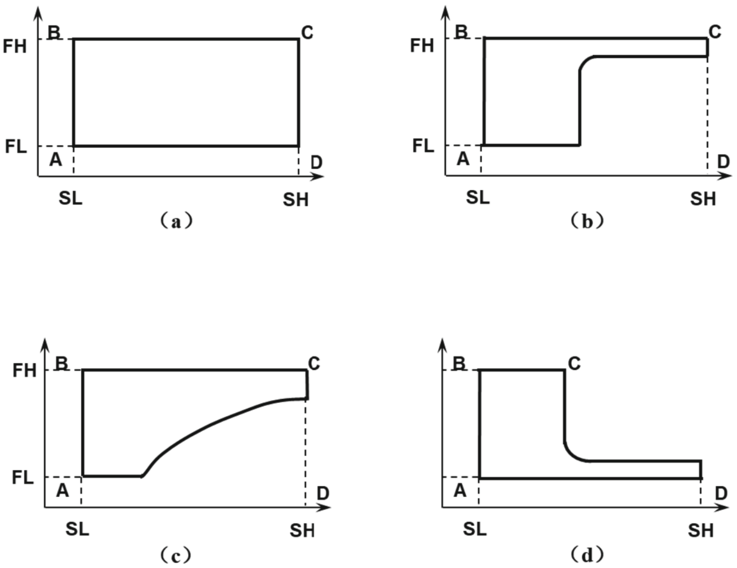


**Fig. 2.** Typical plunger effective stroke. **a**: Normal. **b**: Liquid pound. **c**: Gas effect. **d**: Plunger removal pump

### 2.2 Residual Neural Networks

ResNet, introduced by He et al. [13]. in 2015, is a deep convolutional neural network structure. It was specifically designed to tackle the problems of gradient vanishing and gradient explosion during deep neural network training, enabling more efficient training of deeper networks.

The core concept of ResNet is the incorporation of residual connections, also known as skip connections. These connections enable direct flow of information from shallower layers to deeper layers, preventing the loss or degradation of information within the network. The basic building block of ResNet is the residual block, as depicted in Fig. 3. It consists of two main components: identity mapping and residual mapping.

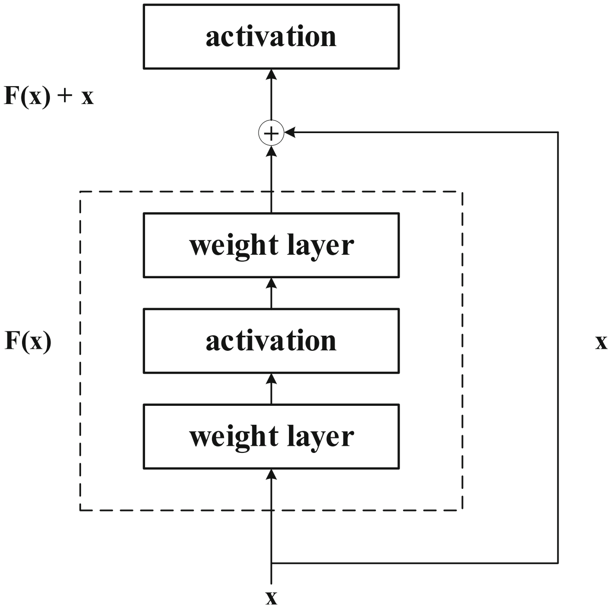


**Fig. 3.** Residual learning: a building block

When the number of channels of the identity mapping $x_i$ is the same as the residual mapping $F(x_i)$, the output of the residual block can be obtained using the following equation:

$$x_{i+1} = x_i + F(x_i, w_i) \tag{5}$$

When the number of channels is different, dimension matching is required by applying a convolutional kernel $W_s$ to adjust the dimensions.:

$$x_{i+1} = W_s \cdot x_i + F(x_i, w_i) \tag{6}$$

### 2.3 Attention Mechanism

The core idea of the Attention mechanism is to simulate the attention mechanism humans employ when processing information. In traditional deep learning models, each input is assigned the same weight and attention, but this is not always the most effective approach. On the contrary, the Attention mechanism allows the model to dynamically adjust attention allocation based on the relevance of the inputs. The calculation formula of attention mechanism is:

$$O = softmax(\frac{QK^T}{\sqrt{L}}).V \tag{7}$$

where $O$ is the output; $Q$ is the input features; $K$ and $V$ the key-value pairs, which are directly derived from the input sequence; $L$ is the input feature length [14].

## 3 Hybrid Model for Production Prediction

### 3.1 Hybrid Model

The hybrid model for production prediction consists of several modules: an input module, a data-driven model, a physics-driven model, and an attention mechanism module. The specific architecture is shown in Fig. 4.

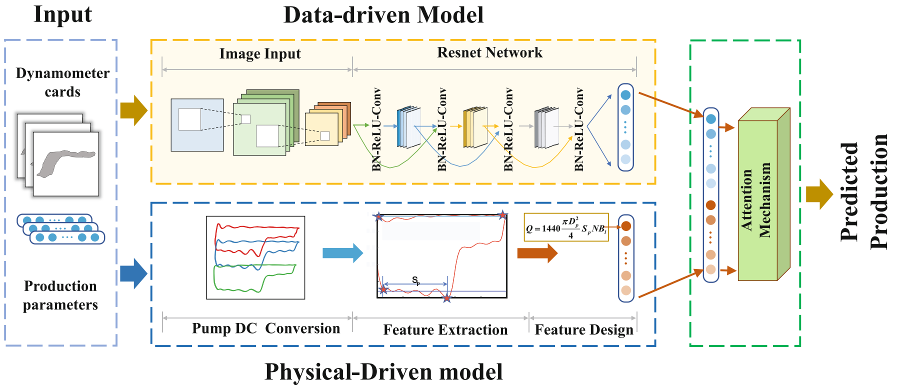


**Fig. 4.** Overview of the proposed Hybrid Model

As shown in Fig. 4, the hybrid model takes as input parameters both the surface dynamometer cards and daily production parameters, such as stroke count, pump diameter, and water cut. The surface dynamometer cards are processed by the data-driven module to extract deep features and obtain a data feature matrix that represents the high-dimensional features.

Simultaneously, the production parameters, along with the surface dynamometer cards, are analyzed by the physics-driven model. This analysis results in a physical feature matrix, which includes conventional sucker rod well production calculations and other physical characteristics.

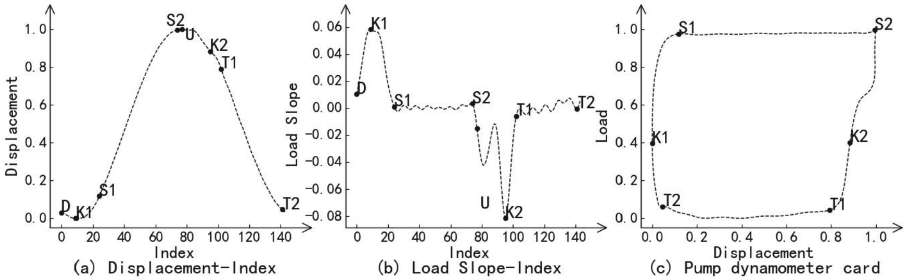
These data and physical feature matrices serve as inputs to the attention mechanism module, where attention weights are dynamically assigned to the outputs of the data-driven and physics-driven models. The attention mechanism evaluates the relevance and importance of the predictions generated by each module, considering the specific task and input conditions.

By combining the deep features extracted from the data-driven module and the physical features obtained from the physics-driven module, the hybrid model aims to leverage the complementary strengths of both approaches. This integration enables a more comprehensive representation of the input parameters, leading to enhanced the accuracy of production predictions in the context of the sucker rod well system.

### 3.2 Physics-Driven Model

The physics-driven model is primarily based on the conventional dynamometer card production measurement technique introduced in Sect. 2.1. As shown in Fig. 4, It begins by mathematically modeling and solving the motion system of the sucker rod well to obtain the pump dynamometer card that represents the downhole pump's motion characteristics.

Subsequently, in feature extraction step, the pump dynamometer card is quantitatively analyzed and computed to identify the switch positions of the traveling valve and the fixed valve. This information is then used in Eq. (4) to calculate the theoretical liquid production rate.



(a) Displacement-Index (b) Load Slope-Index (c) Pump dynamometer card

**Fig. 5.** Feature points extraction. (a) is the displacement curve of the data points. (b) is the slope curve of the normalized load variation of the data points. (c) is the normalized pump dynamometer card.

Additionally, by combining the analysis of displacement curve and load slope curve of pump dynamometer card [15], various physical features are extracted, including geometric slope, average load, valve displacement, and load. The specific steps are follows:

Step 1: In Fig. 5(a), starting from the first data point, search for the first point with a displacement equal to 0, which corresponds to the bottom dead center (D). Also, search for the point with the maximum displacement, which corresponds to the top dead center (U).

Step 2: In Fig. 5(b), identify the point with the maximum slope as $K_1$ and the point with the minimum slope as $K_2$. $K_1$ is located during the upward stroke loading process, while $K_2$ is located during the downward stroke unloading process.

Step 3: In Fig. 5(b), starting from point $K_1$, search forward in the data points for the first point where the slope of the load curve is approximately 0. This point corresponds to the first local maximum between the upward stroke loading process and the top dead center (U), and it is referred to as the fixed valve opening point ($S_1$).

Step 4: In Fig. 5(b), starting from point K1, search forward in the data points until the last point before the top dead center (U) where the slope of the load curve is approximately 0. This point corresponds to the last local maximum between the upward stroke loading process and the top dead center (U), and it is referred to as the fixed valve closing point ($S_2$).

Step 5: In Fig. 5(b), starting from point $K_2$, search forward in the data points for the first point where the slope of the load curve is approximately 0. This point corresponds to the first local minimum between the downward stroke unloading process and the bottom dead center (D), and it is referred to as the traveling valve opening point ($T_1$).

Step 6: In Fig. 5(b), starting from point $K_2$, search forward in the data points until the last point before the bottom dead center (D) where the slope of the load curve is approximately 0. This point corresponds to the last local minimum between the downward stroke unloading process and the bottom dead center (D), and it is referred to as the traveling valve closing point ($T_2$).

Step 7: In Fig. 5(c), record the load values of each data point between the fixed valve opening point (S1) and the fixed valve closing point ($S_2$), and calculate the average load during the upward stroke.

Step 8: In Fig. 5(c), calculate the difference in displacement between the fixed valve opening point ($S_1$) and the fixed valve closing point ($S_2$), which corresponds to the effective stroke during the upward stroke. Also, calculate the difference in displacement between the traveling valve opening point ($T_1$) and the traveling valve closing point ($T_2$), which corresponds to the effective stroke during the downward stroke.

These features, along with the theoretical liquid production rate, are combined to construct the physical feature matrix.

### 3.3 Data-Driven Model

From formulas (1)–(4) and the process of constructing the physical feature matrix, it can be observed that conventional production measurement technique involves numerous assumptions and quantitative analyses. However, during the actual production process of oil wells, various operating conditions and unpredictable dynamometer card deformations can adversely affect the quantitative analysis of the valve switch points, leading to deviations in the calculated effective plunger travel. Therefore, the calculation of production using empirical formulas or mathematical models inevitably introduces certain errors, especially under special operating conditions.

To address this issue, as shown in Fig. 4, this paper adopted a data-driven model to extract deep features from the dynamometer card and utilizes an attention mechanism to effectively integrate the physical and data-driven models.

During the training process of the data model, the dynamometer card $\boldsymbol{X_n}$ is first passed through an image input module that includes convolutional and pooling layers for initial image feature extraction:

$$\mathbf{X_{conv}} = f[conv(\mathbf{X_n} * \mathbf{W}_c) + b] \tag{8}$$

$$\mathbf{X_{map}} = [\max\{\mathbf{X_{conv}}\}] \tag{9}$$

where $\mathbf{X_{conv}}$ is the convolutional layer output; $\mathbf{X_{map}}$ is the pooling layer output; $conv(\cdot)$ stands for the convolution operation; $\mathbf{W_c}$ is the convolution kernel.

Precise prediction of liquid production from dynamometer card images requires accurate extraction of features, specifically the characteristics embodied in the variations of valve switch points and curves during the loading and unloading processes. To overcome the limitations inherent to multi-layer neural networks, like gradient vanishing, a residual neural network consisting of multiple residual blocks is designed to further extract high-dimensional image features.

$$\mathbf{X_L} = \mathbf{X_l} + \sum_{i=l}^{L-1} F(\mathbf{X_i}, \mathbf{W}_i) \tag{10}$$

where $\mathbf{X_L}$ is the characteristic of deep unit L; $\mathbf{X_l}$ is the characteristic of shallow element l; Other symbols have the same meaning as in formula (6).

Then, the feature matrix obtained from the analysis of the physical model is connected to the data model through fully connected layers. Together, these features are fed into the attention mechanism module for the final prediction of oil production.

# 4 Case Study and Results

## 4.1 Dataset

In this study, production data from a certain oilfield in China were selected as an example for experimentation. The sample set consists of 6278 dynamometer cards and corresponding production data from 350 sucker rod wells within a period of 30 days. The dataset was subjected to mathematical and statistical analysis based on different operating conditions, as shown in Table 1.

Upon observing the sample quantities, it can be seen that the largest number of samples corresponds to normal operating conditions, followed by insufficient fluid supply situations. Due to the presence of various types of leakage conditions in pumping unit wells, such as fixed valve leakage, traveling valve leakage, and piston leakage, and the relatively low number of samples for each specific condition, the subcategories related to leakage were merged into one category for statistical analysis.

Based on the distribution of sample production, it can be observed that under normal operating conditions, the average production of wells is the highest, followed by the

**Table 1.** Production statistics under different working conditions.

| | Normal | Insufficient supply | Gas influence | Pump Hitting | Leak | All |
|---|---|---|---|---|---|---|
| Count | 3232 | 2251 | 232 | 261 | 302 | 6278 |
| Mean | 35.94 | 19.91 | 11.12 | 12.74 | 29.96 | 27.98 |
| Min | 3.00 | 0.57 | 1.70 | 0.49 | 10.77 | 0.49 |
| 25% | 21.35 | 9.73 | 5.81 | 8.83 | 17.73 | 14.89 |
| 50% | 29.61 | 15.84 | 11.13 | 14.58 | 29.16 | 22.40 |
| 75% | 45.04 | 23.78 | 13.23 | 16.76 | 39.04 | 35.26 |
| Max | 213.44 | 182.23 | 50.85 | 22.41 | 84.44 | 213.44 |
| Std | 21.34 | 17.24 | 7.85 | 5.17 | 13.55 | 20.78 |

**Table 2.** Model Evaluation Results.

| Model | RMSE | | MAPE(%) | |
|---|---|---|---|---|
| | Training set | Testing set | Training set | Testing set |
| Hybrid Model | **3.16** | **2.82** | **4.31** | **4.33** |
| DModel | 11.35 | 14.27 | 14.34 | 15.26 |
| PModel | 4.12 | 4.45 | 4.87 | 5.01 |
| PDModel | **3.67** | **3.54** | **4.67** | **4.86** |
| SVM | 6.42 | 6.73 | 7.16 | 7.34 |
| XGBoost | 5.51 | 5.82 | 6.13 | 6.76 |
| MLP | 8.76 | 9.54 | 10.43 | 10.12 |

leakage condition. This indicates that most wells experiencing leakage have relatively mild leakage situations and lower leakage volumes. The condition with the lowest average production is gas influence, as in this oilfield, most wells affected by gas experience gas lock phenomena, resulting in minimal liquid production.

## 4.2 Evaluation Metrics

When evaluating regression algorithms, their performance is typically assessed by examining the magnitude of the differences between their predicted results and the true values. The most commonly used evaluation metrics for regression models are the Root Mean Square Error (RMSE) and the Mean Absolute Percentage Error (MAPE).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \tag{10}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \cdot 100\% \tag{11}$$

where $y$ is the true value; $\hat{y}$ is the predicted value from the model. A larger value for both RMSE and MAE indicates a larger difference between the predicted results and the true results of the model, which suggests that the model has lower accuracy in its predictions.

### 4.3 Performance Verification Based on Ablation Study

Performance verification based on ablation study is a crucial step in assessing the effectiveness and contribution of different components or factors within a machine learning or deep learning model. In this section, we will conduct model performance testing and comparisons by employing a hybrid drive model, various ablation models, and conventional benchmark machine learning models.

The ablation models used in the study include the following:

(1) DModel: This model represents a data-driven approach that solely relies on ResNet as the primary component for prediction. It utilizes the deep learning to extract features of dynamometer cards and make predictions.
(2) PModel: In this model, only the physical feature matrix is constructed and fed into the attention mechanism model for predicting production.
(3) PDModel: This model is a modified version of the hybrid drive model, where the attention mechanism is removed.

In addition to the ablation models mentioned earlier, this study also includes several benchmark models for comparison, predicting the daily fluid production from the physical feature matrix.. The benchmark models are as follows:

(1) Support Vector Machine (SVM): SVM works by finding an optimal hyperplane that separates different classes or predicts continuous values based on the data.
(2) XGBoost: XGBoost is a gradient boosting algorithm which combines the power of decision trees and gradient boosting techniques to create an ensemble of weak models that collectively make accurate predictions.
(3) Multilayer Perceptron (MLP): MLP is a type of artificial neural network with multiple layers of interconnected nodes. It is widely used for various machine learning tasks, including regression.

The experiment details are as follows: The dataset was divided into a training set and a test set in a 4:1 ratio, with 4,708 samples in the training set and 1570 samples in the test set. During the training process, each model underwent a random grid search to determine the best-performing model. In the testing phase, both the hybrid drive model and the other six comparative models were evaluated on the test set. The partial fitting performance of the hybrid model on the test set is illustrated in Fig. 6. The comparison of fitting between the training set and the test set is depicted in Fig. 7. The specific results of the ablation study comparison are presented in Table 2.

From Fig. 6 and Fig. 7, it can be observed that the hybrid model demonstrates satisfactory production prediction accuracy in both the training and test sets. However, there

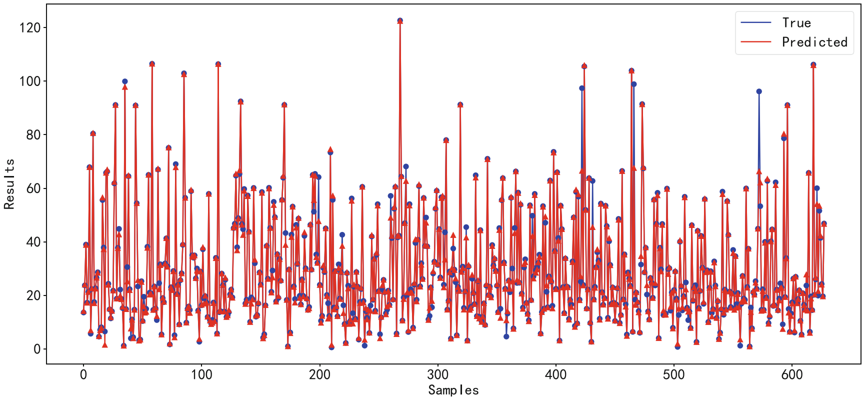


**Fig. 6.** Comparison of actual production and predicted production of hybrid drive model.

are some samples where the model predicts significantly lower production compared to the actual values. Upon further inspection of these wells, it was discovered that besides the model error, some wells were operating in a "gushing with pumping" state, where the surface production rate significantly exceeded the downhole pump's maximum theoretical displacement. This situation deviates from the overall distribution of the oil well sample set and makes it challenging for the model to predict such high production rates accurately. Therefore, the model's prediction accuracy still remains at a high level.

Analyzing the results from Table 2, it can be observed that the hybrid model exhibits the lowest RMSE and MAPE losses, indicating that the proposed model outperforms other conventional production forecasting models. The accuracy of the model on the test set reaches 95.67%, which is at least 2.43% higher than that of the baseline model.

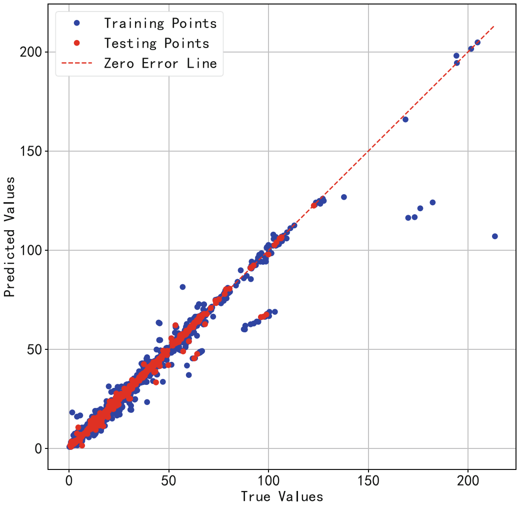


**Fig. 7.** Training and Test set fit plots.

Specifically, PDModel ranks second, indicating that the attention mechanism enables the hybrid model to better capture the weight relationship between the image feature matrix and the physical feature matrix, focusing on the most influential features for production forecasting.

Moreover, compared to the baseline models that solely use the physical feature matrix and exhibit lower accuracy, PDModel leverages Resnet for deep feature extraction from the dynamometer card images, while PModel incorporates the attention mechanism to adapt the internal weights of the physical feature matrix, resulting in significantly improved production prediction accuracy.

It is worth noting that the DModel, which solely uses Resnet for extracting pump dynamometer card features, performs the poorest. This is because it lacks input of important production features specific to oil wells, such as stroke count and pump diameter.

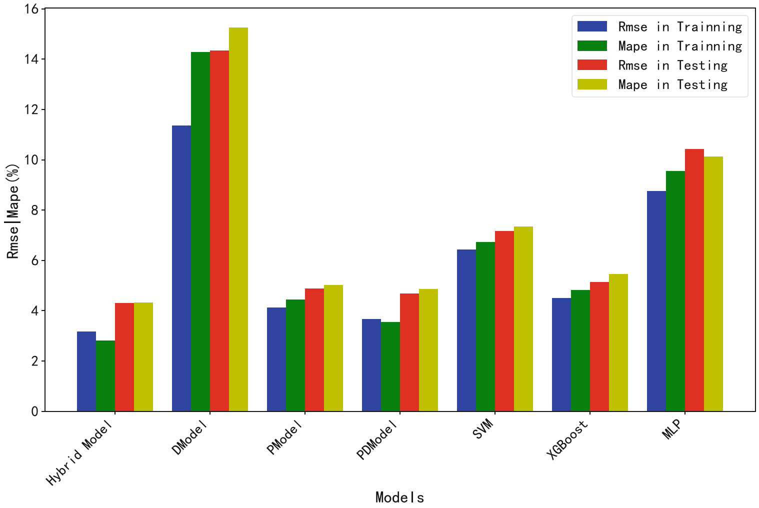


**Fig. 8.** Histogram of model scores.

## 4.4 Performance Verification Based on Different Working Conditions

In this section, the hybrid model and baseline models will be used to predict liquid production in five different operating conditions: Normal, Insufficient supply, Gas influence, Pump Hitting, and Leak. The goal is to analyze and compare the robustness and generalization of the hybrid driving model. Detailed information about the dataset has been presented in Sect. 4.1.

From Table 3, it can be observed that the hybrid driving model exhibits excellent accuracy in the Normal operating condition. Additionally, in the abnormal operating conditions, it maintains an error of less than 10%. Compared to the corresponding optimal baseline models, it achieves an improvement of around 2% in accuracy, demonstrating good generalization and robustness. It is worth noting that although the hybrid model achieves a relatively high average Mape in the Pump Hitting condition, its RMSE is only 1.72. After observing the distribution of liquid production in the sample of Pump Hitting conditions in the dataset, it can be found that the overall liquid production of the oil well under this operating condition is relatively low, with an average value of 12.74 $m^3$/d and a minimum value of only 0.49 $m^3$/d. Therefore, in cases where the sample size is small and the average value is low, even if the RMSE is only 1.72, the relative accuracy of the model prediction will be greatly affected.

Overall, the number of samples for some special operating conditions in the dataset used is relatively small, which is consistent with the uneven nature of oil well operating conditions in the actual production process. Even under these conditions, the hybrid model can still predict oil well fluid production with high accuracy. Therefore, if the sample is equalized through human operation, the accuracy of the hybrid model will be significantly improved under special operating conditions. Alternatively, in future work, it is necessary to consider combining more comprehensive machine learning algorithms and big data processing techniques to reduce the negative impact of sample imbalance on the overall performance of hybrid models.

**Table 3.** Model characteristics under different operating conditions.

| Conditions | Hybrid Model | | Best Baseline Model | |
|---|---|---|---|---|
| | RMSE | MAPE(%) | RMSE | MAPE(%) |
| Normal | 2.35 | 1.97 | 3.37 | 6.05 |
| Insufficient supply | 4.17 | 5.40 | 3.89 | 8.20 |
| Gas influence | 2.76 | 6.31 | 3.04 | 7.93 |
| Pump Hitting | 1.72 | 9.38 | 1.91 | 11.29 |
| Leak | 0.96 | 7.65 | 1.23 | 9.60 |

## 5 Discussion and Conclusion

In this paper, a physical-data hybrid-driven liquid production prediction method based on the attention mechanism is proposed to solve the problem of automatic and accurate measurement of oil well liquid production. This model fuses the weight relationship of the dynamometer card image feature and the physical feature matrix, and realizes the effective combination of features through the attention mechanism. This allows the model to extract key features from multiple perspectives to better understand the relationship between well conditions and fluid production. It provides a powerful tool for oilfield automation management and intelligent development.

In the future work, more machine learning algorithms will be used to solve the problem that the production prediction of sucker rod wells is greatly affected by the type of working conditions, so as to achieve ideal accuracy under special working conditions.

**Acknowledgments.** The project is supported by the National Natural Science Foundation of China under Grant 52274057, 52074340 and 51874335, the Major Scientific and Technological Projects of CNPC under Grant ZD2019-183-008, the Major Scientific and Technological Projects of CNOOC under Grant CCL2022RCPS0397RSN, the Science and Technology Support Plan for Youth Innovation of University in Shandong Province under Grant 2019KJH002, 111 Project under Grant B08028.

## References

1. Ren, T., Sun, C., Sun, W., Kang, X.: The research of metering well yield technology based on dynamometer card. In: Proceedings of the 5th International Conference on Mechanical Engineering, Materials and Energy (5th ICMEME2016). Atlantis Press, Hong Kong, China (2016)
2. Cheng, J., Chen, W., Tao, F., Lin, C.-L.: Industrial IoT in 5G environment towards smart manufacturing. J. Ind. Inf. Integr.Integr. **10**, 10–19 (2018)
3. Mantecon, J.C.: Quantitative interpretation of the surface dynamometer card. Presented at the SPE Asia-Pacific Conference, September 13 (1989)
4. Lyu, X., Ren, X.: An interactive oil well production prediction method for sucker-rod pumps based on dynamometer diagram. In: Proceedings of 2013 2nd International Conference on Measurement, Information and Control, pp. 31–35 (2013)
5. Yin, J.-J., Sun, D., Yang, Y.: Predicting multi-tapered sucker-rod pumping systems with the analytical solution. J. Petrol. Sci. Eng. **197**, 108115 (2021)
6. Lv, X., Wang, H., Liu, Y., Chen, S., Lan, W., Sun, B.: A novel method of output metering with dynamometer card for SRPS under fault conditions. J. Petrol. Sci. Eng. **192**, 107098 (2020)
7. Feng, J., Li, F., Lu, S., Liu, J., Ma, D.: Injurious or noninjurious defect identification from MFL images in pipeline inspection using convolutional neural network. IEEE Trans. Instrum. Meas.Instrum. Meas. **66**(7), 1883–1892 (2017)
8. Ruiz- Serna, M.A., Alzate- Espinosa, G.A., Obando- Montoya, A.F., Álvarez- Zapata, H.D.: Combined artificial intelligence modeling for production forecast in a petroleum production field. CT&F Cienc. Tecnol. Futuro. **9**(1), 27–35 (2019)
9. Pan, S., Wang, J., Zhou, W.: Prediction on production of oil well with attention-CNN-LSTM. J. Phys.: Conf. Ser. **2030**(1), 012038 (2021)
10. Peng, Y., et al.: Deep autoencoder-derived features applied in virtual flow metering for sucker-rod pumping wells. In: Day 1 Tue, October 29, 2019, p. D011S002R003. SPE, Bali, Indonesia (2020)
11. Zhang, K., et al.: Fault diagnosis method for sucker rod well with few shots based on meta-transfer learning. J. Petrol. Sci. Eng. **212**, 110295 (2022)
12. Gibbs, S.G., Neely, A.B.: Computer diagnosis of down-hole conditions in sucker rod pumping wells. J. Petrol. Technol. **18**(01), 91–98 (1966)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015). http://arxiv.org/abs/1512.03385
14. Zhen, Y., Fang, J., Zhao, X., Ge, J., Xiao, Y.: Temporal convolution network based on attention mechanism for well production prediction. J. Petrol. Sci. Eng. **218**, 111043 (2022)
15. Zhang, R., Yin, Y., Xiao, L., Chen, D.: A real-time diagnosis method of reservoir-wellbore-surface conditions in sucker-rod pump wells based on multidata combination analysis. J. Petrol. Sci. Eng. **198**, 108254 (2021)

# Study of Spatial Feature Extraction Methods for Surrogate Models of Numerical Reservoir Simulation

Jin-ding Zhang[1], Kai Zhang[1,2](✉), Li-ming Zhang[1], Pi-yang Liu[2], Wen-hao Fu[1], Wei-long Zhang[1], and Jin-zheng Kang[1]

[1] School of Petroleum Engineering, China University of Petroleum (East China), Qingdao, China
zhangkai@upc.edu.cn

[2] Civil Engineering School, Qingdao University of Technology, Qingdao, China

**Abstract.** Numerical reservoir simulation is an important technology in reservoir production development, but the computational consumption of numerical simulation is a key factor affecting reservoir history matching, production prediction, and optimization. By constructing a computationally fast machine learning model to learn the mapping relationship between reservoir model parameters and production data, a maximum alternative to the numerical simulation process can be achieved to improve the efficiency of reservoir management and decision making. The current surrogate models of reservoir numerical simulation for large spatial variables, including permeability and porosity fields, often extract spatial features by convolutional neural networks and later use recurrent neural networks to learn the time-series relationships of production data. In this work, we study the method using convolutional neural networks to extract spatial parameters of reservoir models and propose a new module to convert the temporal and spatial features of surrogate models. By converting the spatial features extracted by convolution and adapting the input features and dimensions of the recurrent neural network, maximum extraction of spatial feature parameters is achieved. The proposed method was verified on a 3D reservoir model, and the results indicate that the method can enhance the accuracy of the surrogate model.

**Keywords:** Surrogate model · History matching · Spatial feature · Convolutional neural network

## 1 Introduction

In reservoir production and development, numerical simulation technology is a key and effective method to simulate the dynamic process of subsurface reservoirs [1, 2]. By constructing mathematical and physical models, reservoir simulation can assist engineers in understanding the physical parameters and fluid flow patterns of subsurface reservoirs, and developing development plans and managing reservoirs [3]. Reservoir numerical simulation has been widely used in automatic history matching, production forecasting, production optimization, and other processes, and has demonstrated its great advantages [4–8].

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 174–184, 2024.
https://doi.org/10.1007/978-981-97-0272-5_14

However, reservoir numerical simulation involves solving partial differential equations, and the process is computationally expensive. For real reservoir models with millions of grid blocks, a single numerical simulation process can even take several hours. For automatic history matching, reservoir model parameters need to be adjusted many times to fit the historical observations (e.g., oil/water/gas production rate, bottom-hole pressure) [9, 10]. If the observations are matched well, the reservoir model is regarded as the closest to the real reservoir. Then, the calibrated model is reliable for production forecast and optimization. The process always requires thousands of numerical simulations and is even more computationally expensive, which greatly affects the efficiency of the decision-making for oilfields.

In order to speed up the process of reservoir history matching, surrogate modeling was proposed [11–13]. The surrogate model approach is based on machine learning to train a black box model with the samples of the input and output. The model can discover the complex nonlinear relationship between reservoir parameters and the simulation results. Given the input, the surrogate can estimate its corresponding output. In this paper, we are concerned with data-driven surrogates based on artificial neural networks (ANN). The surrogate models can be mainly categorized into online and offline models [14–16]. The key reason for using the online models is the low accuracy of the surrogates (e.g., the radial basis function model, and K-nearest neighboring model), which requires lots of samples in the optimization to retrain the surrogate model and improve its accuracy. Nevertheless, the offline method mainly relies on the performance of deep learning models, which only train the model once with samples and no further training subsequently. With the rapid development of machine learning and neural networks, deep learning methods have a stronger approximation ability for complicated problems and wider applicability [14–19].

For surrogate models of history matching, the inputs are the uncertain parameters of reservoir models and the outputs are well production data. The parameters for reservoir models are often high-dimensional because they are related to the number of grids, which can reach millions. It is still a challenge to extract the features of high-dimensional model parameters and predict their corresponding production data.

In this paper, a feature extraction approach combining convolutional neural networks (CNN) [20] with recurrent neural networks (RNN) [21] is proposed for high-dimensional reservoir model parameters for history matching. The method first extracts reservoir spatial features by the CNN and residual blocks and then the spatial features are input into the RNN to predict the production dynamic data of the reservoir. In order to suit high-dimensional model spatial features, we propose a new spatial and temporal transformation module to retain the spatial features of reservoir model parameters to the maximum extent possible. The proposed method was tested in a 3D reservoir model, and the results show that the method effectively can process the spatial characteristics and enhance the accuracy of the surrogate.

The remainder of the paper is structured as follows. Firstly, the surrogate model based on CNN and RNN is introduced in Sect. 2.1. Then, the proposed transformation module for the spatial and temporal features is presented in Sect. 2.2. After that, a case is used to demonstrate the efficiency of our approach in Sect. 3. Lastly, the conclusions and discussion are given in Sects. 4.

## 2 Methodology

The structure of the surrogate model for history matching is first introduced in this section. The basic modules of the surrogate model include CNN and RNN. Then, the proposed method for extracting the spatial features of the reservoir model is presented.

### 2.1 Surrogate Model Based on CNN and RNN

A surrogate model is a model that replaces a time-consuming numerical simulator. For history matching, e.g., matching production data of oil and water wells, the inputs and outputs of the proxy model are based on the inputs and outputs of the numerical simulator. In this paper, we consider surrogate models for predicting the production data of oil and water wells with time-series characteristics. The architecture of the surrogate mode is mainly referred to [14–16]. The relationship $f$ between the input $m$ and output $d$ using the surrogate model is given by:

$$f : m \rightarrow d \tag{1}$$

where $m \in R^{n_x \times n_y \times n_z \times n_f}$ denotes the model parameters; $d \in R^{n_t \times n_{pf}}$ denotes the production data including injection and production wells; $n_x$, $n_y$, and $n_z$ are the number of grid blocks in the $x$, $y$, and $z$-direction of the model; $n_f$ denotes the type number of model parameters (parameters that rely on grid blocks are considered in this paper, for example, the permeability, porosity, net-to-gross.); $n_{\mathrm{t}}$ denotes the number of timesteps; $n_{pf}$ denotes the type number of production data in a timestep.

The surrogate model includes two main modules, Module 1 based on CNN, and Module 2 based on RNN. CNN is utilized to process the spatial features of reservoir parameters, as depicted in Fig. 1. The model parameters are input into CNN to extract features. In the last layer of the CNN, the spatial features can be obtained using the global average pooling. RNN including the long short term memory (LSTM) [22] and the gated recurrent unit [23] can be used to capture the temporal features of production data, as shown in Fig. 2. The spatial features are repeated to fit time steps of the production data to input into the RNN. After the RNN, the linear layer is added to connect the output of the RNN and the production data. The red dashed box in Fig. 1 and Fig. 2 represents the module for the transformation of the spatial and temporal features. For the surrogate model, this transformation module is very significant because it affects the transfer of information about the spatial characteristics of the reservoir model, which in turn affects the final prediction performance.

### 2.2 Proposed Transformation Module for the Spatial and Temporal Features

In our experiments, we found the transformation module in Fig. 1 and Fig. 2 (marked by the red dashed box) can obtain good results but not the best results for some problems. One of the key reasons is that the final global average pooling of the CNN reduces the parameters while losing some spatial knowledge information of the reservoir model. Thus, we propose a new transformation module to replace the parts in the red dashed box, as shown in Fig. 3.

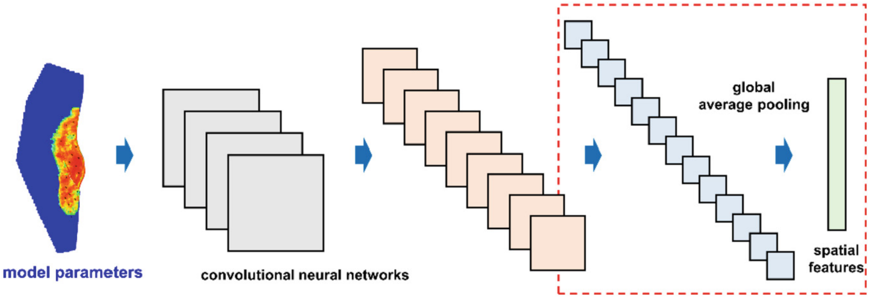


**Fig. 1.** Module 1 for extracting spatial features of the model parameters.

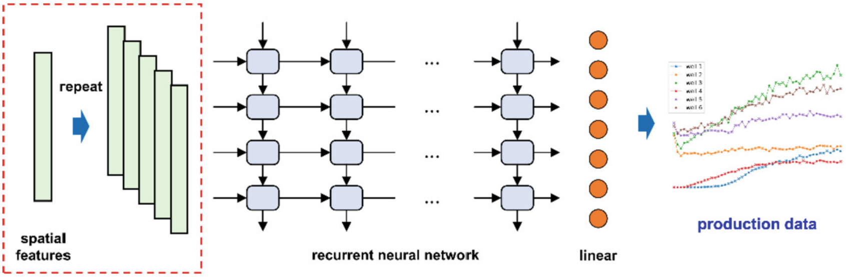


**Fig. 2.** Module 2 for processing temporal features of production data.

An adaptive channel module is constructed to make the size of the output of CNN match the timesteps. The number of the channel $c$ in the adaptive channel module can be calculated by:

$$c = \mathbf{int}(c_{-1}/n_t + 1) \times n_t \tag{2}$$

where **int** denotes rounding the value, and $c_{\text{-}1}$ represents the channel of the last layer for the CNN.

After that, the output of the adaptive channel module is flattened to get the spatial features. The spatial features are then reshaped to equal the number of time steps of the production data. The difference between the proposed adaptive module and the previous surrogate model is whether the spatial feature information is distributed to each time-step feature is the same (after the reshape operation). If the input features are different at each time step, then it is more suitable for RNN training and prediction.

In order to make this machine learning model with stronger approximation capability, it is usually necessary to use more network layers and increase the number of trainable parameters. But as the number of network layers increases, the performance of the surrogate model may even decrease. Thus, to further enhance the effect of the surrogate, the residual block is used in the module, as presented in Fig. 4. The residual block includes the convolutional neural networks, batch normalization [24], and the rectified linear units (ReLU) [25]. The feed-forward of the neural network module is performed through two branches and the input data $x$ is also added behind the final batch normalization layer. For more details, please refer to [16, 26].

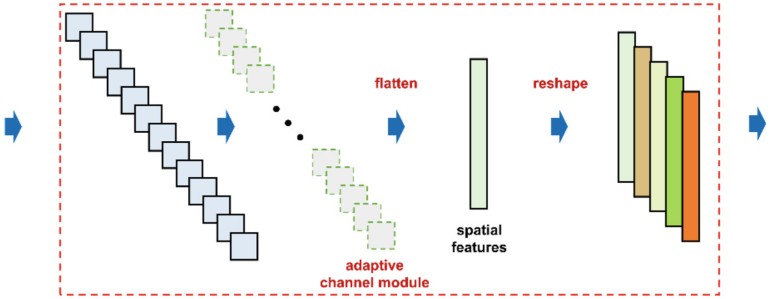


**Fig. 3.** The proposed module for the transformation between the spatial and temporal features.

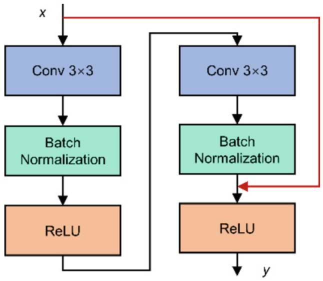


**Fig. 4.** Structure of the residual block in the module.

# 3 Case Study

## 3.1 Reservoir Model

We tested the proposed surrogate model on the Brugge case [27]. This model is a benchmark to analyze and verify the efficiency of waterflooding optimization and history-matching methods. The Brugge model has a grid of 139 × 48 × 9, as presented in Fig. 5. There are 10 injection wells and 20 production wells. The production duration is 10 years, which is divided into 253 timesteps. The uncertain parameters in history matching include permeability, porosity, and net-to-gross thickness ratio. These parameters are the inputs of the surrogate model. The outputs of the surrogate include the bottom-hole pressure of 10 injection wells and the water production rate and oil production rate of 20 production wells (a total of 70 indexes).

A data-driven proxy model is used to predict the production performance of this reservoir. The model has a dimensionality of 30024 for the uncertainty parameters and 17710 for the output production data. The high dimensionality of this problem makes it more difficult for traditional surrogate models, including polynomial regression, kriging methods, radial basis neural networks, etc.

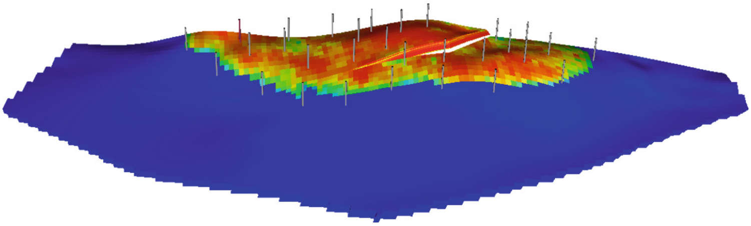


**Fig. 5.** Brugge reservoir model.

### 3.2 Parameter Setting

There are 2000 samples generated to test the surrogate model, 1600 samples are formed as the training set and 200 samples as the validation set, and 200 samples are as the testing set. Figure 6 shows the model parameters in the samples. Each column in Fig. 6 denotes a sample and each row denotes a type of parameter. From top to bottom are net-to-gross, permeability in *x*, *y*, and *z* directions, and porosity. Only the first layer of the simulation model is shown here. Figure 7 shows the range of production data of the samples. The first 7590 data are the bottom-hole pressures of the production and injection wells, the 7590–12650 data are the oil production rates of the production wells, and the 12650–17710 data are the water production rates of the production wells. For the training of the surrogate, the training epochs are 100. The learning rate is 0.003 and is set to decrease adaptively. The batch size is set to 16.

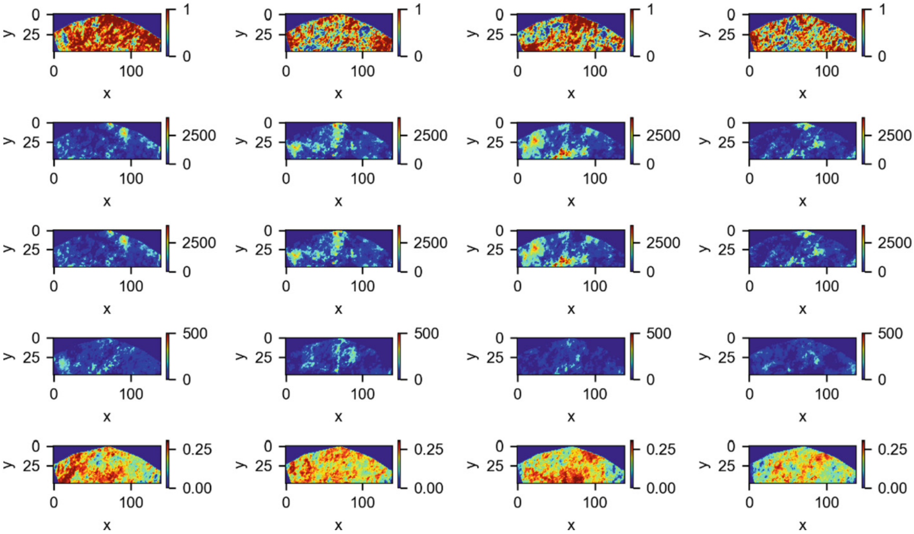


**Fig. 6.** Model parameters of the training samples. Each column denotes a sample and each row denotes a type of parameter. From top to bottom are net-to-gross, permeability in x, y, and z directions, and porosity. Only parameters of the first layer of the model are given.

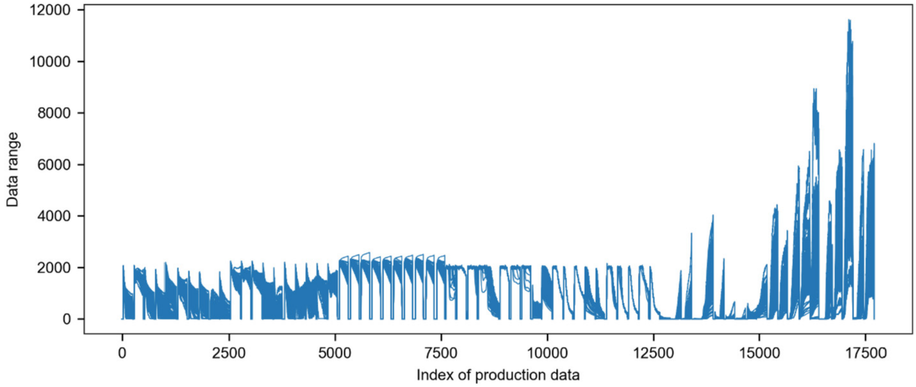


**Fig. 7.** Range of production data in the training samples (a total of 17710 data (253 × 70)).

### 3.3 Results

We compared the proposed surrogate model (CNN-RES-TS) with the surrogate with only the CNN (CNN) and the surrogate with the CNN and residual block (CNN-RES). We used the means square error (*MSE*) and coefficient of determination ($R^2$) to measure the prediction error for the three surrogate models. The *MSE* can be represented by:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{3}$$

where $N$ refers to the number of samples, $y_i$ indicates the surrogate prediction and $\hat{y}_i$ represented sample data.

The $R^2$ can be represented as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y}_i)^2} \tag{4}$$

where $\overline{y}_i$ represents the sample.

Figure 8 presents the *MSE* of the three surrogates in the training, validation, and testing set. Figure 9 shows the coefficient of determination ($R^2$) for three methods. The smaller the *MSE* and the larger the $R^2$ indicate the better the surrogate model. The results show that CNN-RES-TS has the fastest convergence for both MSE and R$^2$ and the final value is the best. Faster convergence of the surrogate model means less training time, so the time consumption for building the surrogate model can be greatly alleviated.

Figures 10, 11, and 12 show the prediction results in the testing set for three methods. Comparing CNN and CNN-RES, CNN-RES obtained better prediction results for the production data index 15000–17710. This indicates that the residual block can enhance the generalization of the surrogate. Comparing CNN-RES and CNN-RES-TS, CNN-RES-TS can obtain a better prediction (especially for the production data index 15000–17710). This indicates that the proposed transformation module can better extract the spatial features of reservoir model parameters, and thus enhance the prediction accuracy of the surrogate.

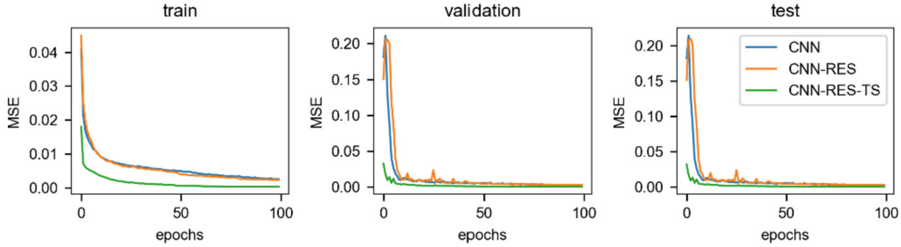


**Fig. 8.** Mean square error (*MSE*) for the training set, validation set, and test set.

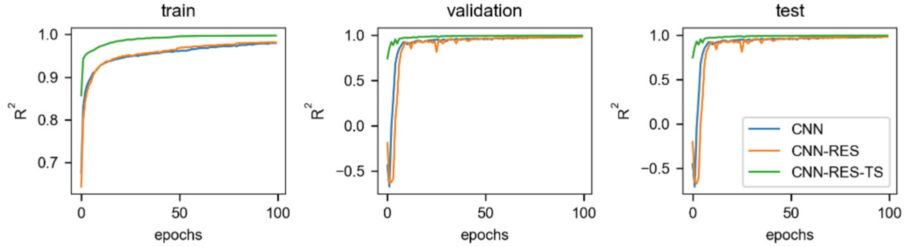


**Fig. 9.** Coefficient of determination ($R^2$) for the training set, validation set, and test set.

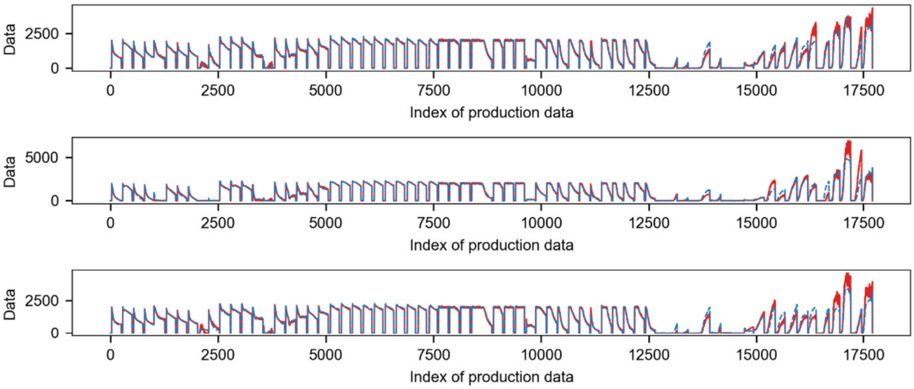


**Fig. 10.** Comparison between the surrogate prediction using CNN and the sample. The red line denotes the sample and the blue dashed line refers to the prediction. Three random samples in the test set are shown.

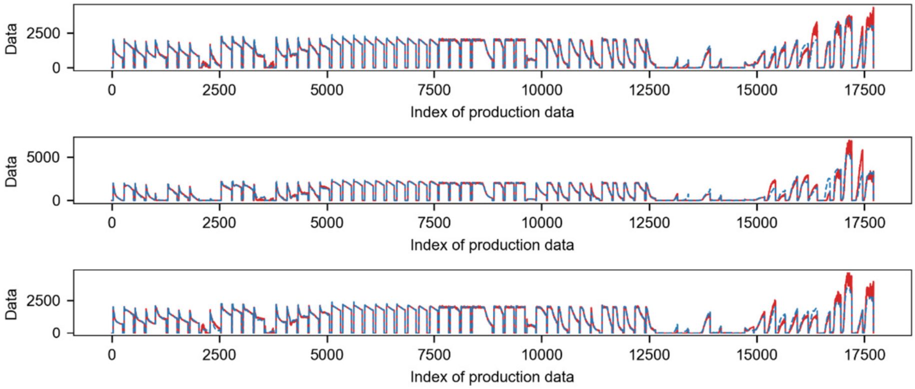


**Fig. 11.** Comparison between the surrogate prediction using CNN-RES and the sample. The red line denotes the sample and the blue dashed line refers to the prediction. Three random samples in the test set are shown.

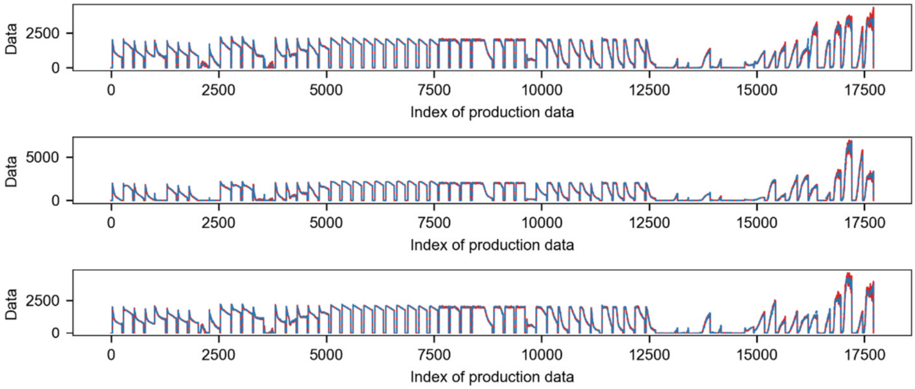


**Fig. 12.** Comparison between the surrogate prediction using CNN-RES-TS and the sample. The red line denotes the sample and the blue dashed line refers to the prediction. Three random samples in the test set are shown.

## 4 Conclusions and Discussion

In this work, we present a new spatial feature extraction module for reservoir model parameters for the surrogate model. The method contains an adaptive channel module and the corresponding spatial feature transformation method. The method was tested and analyzed on a 3D reservoir model, and the results indicate that the proposed feature extraction approach can enhance the prediction accuracy of the surrogate model and provide a reference for the research of surrogate modeling for reservoir numerical simulation. The module proposed in this paper is adaptive and has no parameters that can be set, thus making it easier to build agent models. The main reason for the surrogate model approach, which is still not widely applied in actual oil fields, is that the surrogate model approach requires engineers with a high level of theoretical approaches to machine learning and neural networks. Also, the complexity of production data and reservoir properties can affect the effectiveness of the surrogate model.

In the current research on surrogate models of reservoir numerical simulation, there are many types of surrogate models. It is still a challenge to determine a suitable surrogate model for a specific problem and to explain the mechanism behind them. The parameters of the surrogate model need to be adjusted many times to achieve a satisfying result. This process is very tedious and time-consuming and requires specialized domain knowledge in machine learning, which is still challenging for some reservoir engineers. In terms of surrogate models applied to problems of reservoirs, future research directions include interpretability, hyper-parameter optimization, and automatic design of architectures for surrogate models.

**Acknowledgments.** This work is supported by the National Natural Science Foundation of China under Grant 52274057, 52074340 and 51874335, the Major Scientific and Technological Projects of CNPC under Grant ZD2019-183-008, the Major Scientific and Technological Projects of CNOOC under Grant CCL2022RCPS0397RSN, the Science and Technology Support Plan for Youth Innovation of University in Shandong Province under Grant 2019KJH002, 111 Project under Grant B08028.

## References

1. Peaceman, D.W.: Fundamentals of numerical reservoir simulation. Elsevier (2000)
2. Oliver, D.S., Chen, Y.: Recent progress on reservoir history matching: a review. Comput. Geosci.. Geosci. **15**(1), 185–221 (2010)
3. Ertekin, T., Abou-Kassem, J.H., King, G.R.: Basic applied reservoir simulation (2001)
4. Gilman, J.R., Ozgen, C.: Reservoir simulation: history matching and forecasting. Society of Petroleum Engineers Richardson (2013)
5. Gu, Y., Oliver, D.S.: History matching of the PUNQ-S3 reservoir model using the ensemble Kalman filter. SPE J. **10**(02), 217–224 (2005)
6. Zhang, K., Zhang, J.D., Ma, X.P., et al.: History matching of naturally fractured reservoirs using a deep sparse autoencoder. SPE J. **26**(4), 1700–1721 (2021)
7. Zhong, C., Zhang, K., Xue, X., et al.: Historical window-enhanced transfer gaussian process for production optimization. SPE J. **27**(05), 2895–2912 (2022)
8. Sun, W.Y., Hui, M.H., Durlofsky, L.J.: Production forecasting and uncertainty quantification for naturally fractured reservoirs using a new data-space inversion procedure. Comput. Geosci.. Geosci. **21**(5–6), 1443–1458 (2017)
9. van Leeuwen, P.J., Evensen, G.: Data assimilation and inverse methods in terms of a probabilistic formulation. Mon. Weather Rev. **124**(12), 2898–2913 (1996)
10. Oliver, D.S., Reynolds, A.C., Liu, N.: Inverse Theory for Petroleum Reservoir Characterization and History Matching. Cambridge University Press, Cambridge (2018)
11. Mohaghegh, S.D.: Reservoir simulation and modeling based on artificial intelligence and data mining (AI&DM). J. Natural Gas Sci. Eng. **3**(6), 697–705 (2011)
12. Oladyshkin, S., Class, H., Nowak, W.: Bayesian updating via bootstrap filtering combined with data-driven polynomial chaos expansions: methodology and application to history matching for carbon dioxide storage in geological formations. Comput. Geosci.. Geosci. **17**(4), 671–687 (2013)

13. de Lira, J.D., Willmersdorf, R.B., Afonso, S.M.B., et al.: Automatic history matching considering surrogate-based optimization and Karhunen-Loève expansions. J. Braz. Soc. Mech. Sci. Eng. **36**(4), 919–928 (2014)
14. Ma, X.P., Zhang, K., Wang, J., et al.: An efficient spatial-temporal convolution recurrent neural network surrogate model for history matching. SPE J. **27**(2), 1160–1175 (2022)
15. Ma, X., Zhang, K., Zhao, H., et al.: A vector-to-sequence based multilayer recurrent network surrogate model for history matching of large-scale reservoir. J. Petroleum Sci. Eng., 110548 (2022)
16. Ma, X., Zhang, K., Zhang, J., et al.: A novel hybrid recurrent convolutional network for surrogate modeling of history matching and uncertainty quantification. J. Petrol. Sci. Eng. **210**, 110109 (2022)
17. Xiao, C., Lin, H.-X., Leeuwenburgh, O., et al.: Surrogate-assisted inversion for large-scale history matching: Comparative study between projection-based reduced-order modeling and deep neural network. J. Petroleum Sci. Eng., 208 (2022)
18. Jiang, S., Durlofsky, L.J.: Use of multifidelity training data and transfer learning for efficient construction of subsurface flow surrogate models. arXiv preprint arXiv:220411138 (2022)
19. Tang, M., Liu, Y.M., Durlofsky, L.J.: Deep-learning-based surrogate flow modeling and geological parameterization for data assimilation in 3D subsurface flow. Comput. Methods Appl. Mech. Eng., 376 (2021)
20. Gu, J., Wang, Z., Kuen, J., et al.: Recent advances in convolutional neural networks. Pattern Recogn.Recogn. **77**, 354–377 (2018)
21. Medsker, L.R., Jain, L.: Recurrent neural networks. Des. Appl. **5**, 64–67 (2001)
22. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput.Comput. **9**(8), 1735–1780 (1997)
23. Cho, K., Van Merriënboer, B., Gulcehre, C., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:14061078 (2014)
24. Bjorck, N., Gomes, C.P., Selman, B., et al.: Understanding batch normalization. Advances in neural information processing systems, 31 (2018)
25. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the Proceedings of the 27th International Conference on Machine Learning (ICML-10), F (2010)
26. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition; proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
27. Peters, E., Chen, Y., Leeuwenburgh, O., et al.: Extended brugge benchmark case for history matching and water flooding optimization. Comput. Geosci.. Geosci. **50**, 16–24 (2013)

# Optimized Drilling Status Recognition for Oil Drilling Using Artificial Intelligence: Empirical Research and Methodology

Xin-yi Yang[1,2](✉), Meng Cui[1,2], Yan-long Zhang[1,2], Ling-zhi Jing[1,2], Yong Ji[1,2], and Xiao-yan Shi[1,2]

[1] CNPC Engineering Technology Research&D Company Limited, Beijing, China
yangxydr@cnpc.com.cn

[2] National Engineering Research Center of Oil & Gas Drilling and Completion Technology, Beijing, China

**Abstract.** This study focuses on the application of modern artificial intelligence (AI) techniques to improve the accuracy of drilling status recognition in the oil drilling industry, with the aim of enhancing safety and efficiency. To address the limitations of existing research in evaluation methods and practical applications, we constructed a unified drilling status dataset, introduced a more scientific evaluation criterion, the F1 score, and conducted a comprehensive evaluation and improvement of existing oil drilling status recognition methods. This paper provides an in-depth analysis of various drilling status characteristics and explores the applicability and limitations of different AI algorithms in drilling status recognition. Based on these findings, we propose optimized general drilling status recognition algorithms and validate the performances in real drilling status. Our research offers valuable insights and guidance for future oil drilling status recognition studies and is expected to promote safer and more efficient development in the oil drilling industry.

**Keywords:** Drilling · Drilling status Recognition · Dataset · Machine Learning · Deep Learning

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20–22 September 2023.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 185–193, 2024.
https://doi.org/10.1007/978-981-97-0272-5_15

# 1 Introduction

With the rapid development of modern artificial intelligence technology, the recognition of oil drilling status has become a hot topic of research. To ensure the safety and efficiency of the drilling process, this paper aims to explore the use of artificial intelligence algorithms, such as machine learning and deep learning methods, to provide new solutions for the recognition of oil drilling status. Despite many recent studies attempting to use artificial intelligence technologies like decision trees [1], support vector machines [2], and deep learning for condition recognition, these studies have certain limitations in evaluation methods and applications. This paper aims to address these limitations, proposing more scientific evaluation criteria and building a unified drilling status dataset.

Firstly, this paper will introduce a new drilling status dataset, which covers various types of status information, providing a solid foundation for the application of artificial intelligence algorithms. Drawing on the classic ImageNet [3] dataset in the field of computer vision, we hope to provide a shared, standardized data foundation for researchers in the field of oil drilling status recognition through the construction of this dataset, thereby promoting algorithm innovation and development.

Secondly, to address the limitations of existing research in evaluation methods, this paper will use a more scientific evaluation standard, namely the F1 score [4]. The F1 score combines precision and recall, enabling a more comprehensive evaluation of model performance. Comparing various algorithms on a unified drilling status dataset, we will be able to gain a deeper understanding of the strengths and weaknesses of various methods, providing more valuable guidance and insights for the field of oil drilling status recognition.

Based on a unified dataset and evaluation standards, this paper will comprehensively assess and improve existing oil drilling status recognition methods. After conducting a detailed empirical analysis of various algorithms, this paper will propose an optimized algorithm for oil drilling status recognition, aiming to achieve high-precision condition prediction, thus enhancing the safety of the drilling process.

In conclusion, the work of this paper will provide a powerful inspiration and guidance for the future development of oil drilling status recognition research. We believe that with the continuous innovation and application of artificial intelligence technology, the field of oil drilling will welcome a safer and more efficient development. By analyzing the characteristics of various conditions and designing corresponding artificial intelligence algorithms based on these characteristics, this paper will help to further advance the research on oil drilling status recognition, providing more effective and reliable solutions for future practical applications.

# 2 Drilling Status Dataset

## 2.1 Collection of Drilling Status

The dataset is collected from real logging data in the Engineering Intelligent Support Center (EISC) system, downloaded through the EISC data lake, and provided by the Xinjiang Oilfield Engineering Institute and the Junggar Project Department in three ways: collecting design documents, logs, well histories, logging, well testing and other data. Data from 23 completed wells in different regions was collected, and the DDR drilling status intelligent recognition system was used to complete the structuring and standardization of the logs; the DDR accident complexity intelligent recognition system was used to identify and construct a complex accident ledger. The data were categorized according to 9 normal drilling statuses including drilling, circulation, reaming, and casing, and calibrated by specialized experts. A total of 224,781 status data records were processed, with 153,776 valid data records (excluding status data with empty features in any field). To ensure the accuracy of the status labels, the data annotation was performed by five field experts, and the final drilling status label was determined by majority vote. This dataset can be used by researchers to study and develop intelligent engineering operation support systems.

## 2.2 Analysis of Drilling Status Dataset

This drilling dataset includes nine drilling statuses: composite drilling, casing running, back reaming, directional drilling, drilling down, circulation, single joint connection, pulling out of hole, and. The specific statistics for each drilling status are shown in Fig. 1: among them, composite drilling is the most common, and reaming data is relatively less.

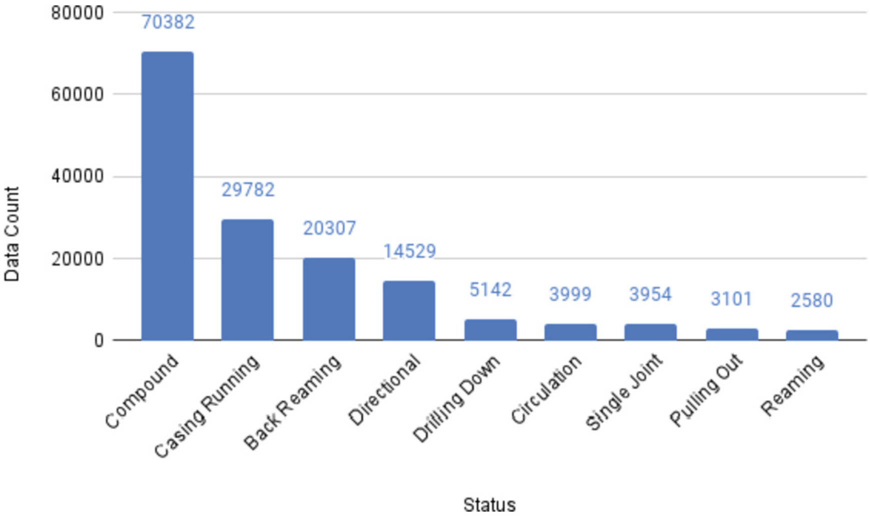


**Fig. 1.** Statistics of different drilling status in the drilling data set

In all drilling status data, there are 17 key features (units in parentheses): torque (kN·m), total pit volume (m$^3$), weight on bit (kN), inlet flow rate (L/s), rotary table speed (rpm), outlet flow rate (L/s), delayed drilling depth (m), standpipe pressure (MPa), well depth (m), number one pump stroke (spm), drill bit position (m), outlet flow rate percentage (%), number two pump stroke (spm), number three pump stroke (spm), hook load (kN), hook height (m), and casing pressure (MPa). These features, which originate from historical manual drilling status judgments, are of crucial value for drilling status recognition. This study will use these features to apply artificial intelligence algorithms to predict drilling status categories, aiming to improve recognition accuracy and practicality.

## 3 Drilling Status Recognition Tasks and Experiments

In the previous section, we introduced the drilling status dataset. In this section, we introduce the task of drilling status recognition into the field of machine learning, thus achieving more efficient and accurate recognition. For this task, we will outline the basic processes involved and describe the core steps of machine learning in handling classification tasks. This paper will focus on using the proposed dataset to evaluate different machine learning algorithms, analyze, and propose future research directions.

### 3.1 Overview of Machine Learning Algorithm Development

Drilling status recognition, as a classification task, aims to use machine learning algorithms to automatically recognize different drilling statuses. Based on this goal, we can divide the entire processing flow into the following key steps, as shown in Fig. 2:

1. Data Preparation: First, the drilling status dataset needs to be preprocessed, including data cleaning, handling missing values, outlier processing, and feature engineering, etc., to ensure data quality and usability.
2. Feature Selection: After data preprocessing, it is necessary to determine the most representative and discriminative features for the specific classification task through feature selection techniques to improve the performance of the classification model.
3. Model Selection and Training: Next, according to task requirements and data characteristics, select an appropriate machine learning algorithm, and use the training dataset to train the model to learn the association between features and drilling status categories.
4. Model Evaluation and Optimization: After model training, predict the test dataset to evaluate the model's performance. If the evaluation results are unsatisfactory, the model can be adjusted and optimized to improve classification accuracy.
5. Application and Deployment: After the above steps, when a classification model that meets the requirements is obtained, it can be deployed in actual drilling scenarios to achieve automatic recognition and monitoring of drilling statuses.

Based on the above process, this paper summarizes and organizes the drilling status recognition task. In this task, our goal is to predict the corresponding drilling status category based on the input drilling status data. Specifically, the input data includes a series of key features during the drilling process, such as torque, total pit volume, weight on bit, etc. The model generates the corresponding drilling status category as output by analyzing these features.

### 3.2 Drilling Status Recognition Algorithm

The purpose of this paper is to explore the performance of different machine learning algorithms in the task of drilling status recognition. Therefore, this study will select

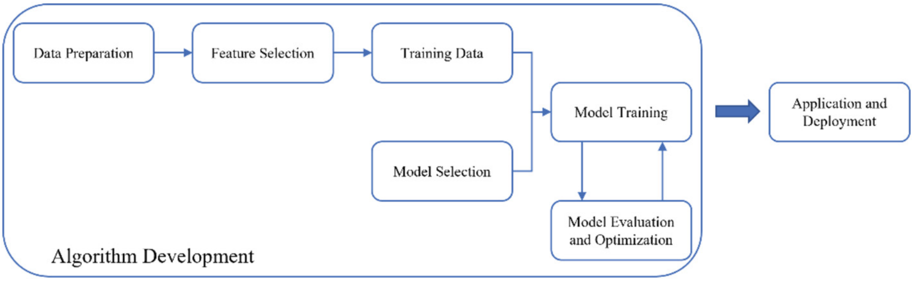


**Fig. 2.** Diagram of machine learning algorithm development for drilling status recognition

various commonly used multi-classification machine learning algorithms for experimentation, including: Logistic Regression (LR) [5], Support Vector Machine (SVM) [6], K-Nearest Neighbors (KNN) [7], Decision Tree Classifier (DTree) [8], Random Forest Classifier (RTree) [9], Multilayer Perceptron (MLP) [9], Gaussian Naive Bayes (GauNB) [9], AdaBoost Classifier (AdaB) [10], Gradient Boosting Classifier (GradB) [11].

In order to further improve the performance of status prediction, this paper introduces ensemble learning methods [12]. Ensemble learning is a strategy of combining multiple weak classifiers to form a strong classifier. In this research, we select models that have performed well in previous experiments as sub-models and construct a Voting Classifier to implement ensemble learning. The core idea of ensemble learning is to make the final prediction more stable and reliable by synthesizing the prediction results of multiple models. Ensemble learning has the following advantages: it reduces the risk of overfitting; improves prediction accuracy; enhances model stability; and handles diverse data.

In this paper, we construct a voting classifier by integrating multiple well-performing sub-models into a powerful classifier. This method is expected to improve the predictive performance of the drilling status recognition task, providing more reliable status recognition results for practical applications.

First, we need to preprocess the data. As the units among different features in the status data are different, we need to normalize the data before training. After normalization, the data will be in a unified scale range, which will help to improve the training effect and performance of the model.

Next, we will use the processed data to train various algorithms and evaluate the training results. In past research, the evaluation indicator usually used was accuracy. However, accuracy does not fully reflect the performance of the model in classification tasks. Therefore, this paper introduces the F1 score as the evaluation criterion.

The F1 score is the harmonic mean of Precision and Recall. Compared with accuracy, the F1 score has the following advantages: first, the F1 score takes into account both the precision and recall of the model, which makes the model have better evaluation effect when dealing with imbalanced datasets; second, the F1 score can calculate evaluation indicators for each category separately in multi-category classification problems, and then give an overall performance evaluation. In summary, the F1 score is a more comprehensive and robust evaluation indicator, which helps to understand the performance of different machine learning algorithms in the drilling status recognition task.

### 3.3 Experiment Analysis

**Table 1.** Prediction results for different machine learning algorithms

| Algorithm | F1 Score % (↑) | Accuracy % (↑) |
|---|---|---|
| LR | 96.2 | 96.1 |
| SVM | 97.2 | 97.2 |
| KNN | 97.6 | 97.7 |
| DTree | 98.4 | 98.4 |
| RTree | 99.0 | 99.0 |
| MLP | 97.5 | 97.5 |
| GauNB | 72.9 | 70.6 |
| AdaB | 62.3 | 67.2 |
| GradB | 98.8 | 98.8 |
| Voting Classifier | 98.1 | 98.1 |

**Which Algorithm Performs Better?**

In the experiments of different machine learning algorithms in the drilling status recognition task, we obtained the F1 scores and accuracy results as shown in Table 1. The analysis is as follows:

Random Forest and Gradient Boosting Classifier performed outstandingly in this experiment, with high F1 scores and accuracy, indicating that these two algorithms have strong predictive ability when dealing with the drilling status recognition task. The performance of Decision Tree is also relatively good, with high F1 scores and accuracy, which can be used as an alternative plan for further optimization and adjustment. Logistic Regression, Support Vector Machine, and K-Nearest Neighbors algorithms performed moderately. Although they may not meet the prediction requirements in this experiment, they may still have certain application value in specific scenarios. The performance of Multilayer Perceptron is close to K-Nearest Neighbors, but slightly inferior to Decision Tree, Random Forest, and Gradient Boosting Classifier. In practical applications, you can try to adjust its parameters to improve prediction performance. Gaussian Naive Bayes and AdaBoost classifiers performed poorly in this experiment, with low F1 scores and accuracy. In the drilling status recognition task, these two algorithms may not be the best choices.

**Is Ensemble Learning Useful?**

The Voting Classifier, as a method of ensemble learning, performed well in the experiment, with both F1 scores and accuracy reaching 98.1%. Although in this experiment, the performance of the Voting Classifier was slightly lower than that of the Random Forest and Gradient Boosting Classifier, it still demonstrated significant predictive capability.

The advantage of ensemble learning methods is that they integrate the prediction results of multiple sub-models, reducing the risk of overfitting of a single model, thereby enhancing the generalization ability of the model. The predictive performance of the Voting Classifier is influenced by the performance of multiple sub-models, so in practical applications, attempts can be made to optimize and adjust the sub-models to further enhance the predictive capability of the Voting Classifier.

In summary, although the performance of the Voting Classifier in this experiment was slightly lower than that of the Random Forest and Gradient Boosting Classifier, as an ensemble learning method, it still demonstrated high predictive performance. Future research could consider further optimization and adjustment of the sub-models of the Voting Classifier, to further improve the accuracy and practicality of drilling status recognition.

**Does Different Features Have Different Impacts on the Model?**

Indeed, feature selection has a significant impact on the performance of the model. By selecting appropriate features, the complexity of the model can be reduced, computational costs can be minimized, and prediction accuracy can be improved. Therefore, feature selection can be an important direction for future research. This article mainly presents a dataset suitable for various condition predictions, proposes a condition prediction task, and tests the performance of commonly used machine learning algorithms on the proposed dataset and tasks, so no special operations for feature selection were conducted. Subsequent research can explore how to select features that are more suitable for predicting all conditions. Furthermore, for different machine learning models, researchers can also try to carry out targeted feature selection to maximize the advantages of each model and further improve prediction performance.

In summary, feature selection is of significant importance in the task of drilling status recognition. Future research can explore from multiple perspectives how to choose more representative features to improve the predictive performance and practicality of the model.

## 4 Future Direction

While the prediction accuracy has already reached about 99%, there are still some research directions worth exploring in the field of drilling condition recognition:

1. Feature engineering: Although the existing features have achieved good prediction results, the feature set can still be optimized to enhance the model's generalizability through further feature engineering, such as feature selection, dimensionality reduction, and feature construction.

2. Model fusion: Try to fuse different types of models, such as stacking, Bagging, and Boosting methods, to enhance the model's stability and generalization performance.
3. Online learning and incremental learning: Drilling condition data may change over time. Researching online and incremental learning methods can enable the model to continuously update and optimize on new data, improving prediction capabilities.
4. Anomaly detection and handling: Anomalies may occur during the drilling process, and these anomalies may affect the prediction performance of the model. Researching anomaly detection and handling methods can enhance the model's robustness when facing abnormal data.
5. Interpretability research: Improve the interpretability of the model, helping engineers understand the reasons for the model's predictions, thus providing more targeted suggestions for drilling operations.

By exploring these research directions, the field of drilling condition recognition will continue to develop in the future, providing higher quality prediction and decision support for the drilling industry.

## 5 Conclusion

This paper introduces a brand-new drilling condition dataset and standardizes the task of condition recognition prediction. Based on the proposed dataset and tasks, we evaluated a variety of different machine learning algorithms and conducted a detailed analysis of the prediction performance of each algorithm. This research result provides a benchmark for subsequent researchers to facilitate more in-depth discussions in the field of condition recognition.

Through experiments and analysis of different machine learning algorithms, we revealed the strengths and weaknesses of each algorithm in the task of drilling condition recognition. In addition, we introduced ensemble learning methods and improved prediction performance by combining multiple excellent sub-models into a voting classifier.

This research not only provides a new data foundation and prediction standard for the task of drilling condition recognition but also provides useful insights for researchers in related fields. Future research can continue to explore more advanced machine learning algorithms and optimization techniques based on this paper, thus achieving more significant results in the field of drilling condition recognition. We hope this research can provide strong support for actual drilling operations and contribute to improving drilling efficiency and safety.

## References

1. Liu, S., Cao, X.: Intelligent recognition method of drilling conditions based on decision trees. New Industrialization **12**(1), 28–30 (2022)
2. Zhang, F., Cui, Y., Yu, C., Zhang, T., Chen, J., Yan, H.: Current situation and development of drilling condition recognition technology based on machine learning. J. Yangtze Univ. (Nat. Sci. Edn.). https://doi.org/10.16772/j.cnki.1673-1409.20230302.001

3. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Li, F.-F.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
4. Sasaki, Y.: The truth of the f-measure (2007). https://www.cs.odu.edu/mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07. Accessed 26 May 2021
5. LaValley, M.P.: Logistic regression. Circulation **117**(18), 2395–2399 (2008)
6. Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B.: Support vector machines. IEEE Intell. Syst. Appl. **13**(4), 18–28 (1998)
7. Peterson, L.E.: K-nearest neighbor. Scholarpedia **4**(2), 1883 (2009)
8. Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A., Brown, S.D.: An introduction to decision tree modeling. J. Chemometr. Soc. **18**(6), 275–285 (2004)
9. Biau, G., Scornet, E.: A random forest guided tour. TEST **25**, 197–227 (2016). https://doi.org/10.1007/s11749-016-0481-7
10. Noriega, L.: Multilayer perceptron tutorial, vol. 4, no. 5. School of Computing. Staffordshire University (2005)
11. Leung, K.M.: Naive Bayesian classifier, pp. 123–156. Polytechnic University Department of Computer Science/Finance and Risk Engineering (2007)
12. Schapire, R.E.: Explaining AdaBoost. In: Schölkopf, B., Luo, Z., Vovk, V. (eds.) Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik, pp. 37–52. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41136-6_5
13. Natekin, A., Knoll, A.: Gradient boosting machines a tutorial. Front. Neurorobot. **7**, 21 (2013)
14. Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q.: A survey on ensemble learning. Front. Comput. Sci. **14**, 241–258 (2020)

# Analysis and Recommendation of Frequent Patterns of Long-Life Pumping Wells Based on Data Mining

Zhong-hui Zhang(✉)

Petroleum Engineering Technology Research Institute of Sinopec Shengli Oilfield Company, Dongying, China
zhangzhonghui.slyt@sinopec.com

**Abstract.** The theoretical regulation of production parameters in oil production engineering plays a significant role in the management of beam pumps. However, it falls short in identifying the inherent relationships among historical production data, thus failing to address the problem at its core. Valuable information can be extracted from historical well experiences through data mining techniques, offering new insights for adjusting production measures. To achieve this objective, an analysis is conducted to explore the factors and patterns influencing the exemption period of oil wells. Various methods, including expert experience and correlation analysis, are employed to process and selectively identify relevant features. Drawing upon the principles of oil production engineering and leveraging advanced big data processing techniques, these features are encoded to construct a comprehensive sample set that represents long-life wells. Subsequently, association rule mining is applied to uncover frequent patterns exhibited by these long-life wells. By setting a minimum support threshold of 0.01, the mining process encompasses a substantial dataset comprising over 1700 wells, leading to the discovery of more than 100 meaningful association rules. These rules are further prioritized and visualized based on their lift values, providing valuable insights into the experiential knowledge base related to effective measures for long-life well patterns. Consequently, this knowledge base becomes an invaluable asset, offering support for informed decision-making in terms of production parameter control and aiding in the development of scientifically guided production strategies.

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20–22 September 2023.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 194–204, 2024.
https://doi.org/10.1007/978-981-97-0272-5_16

**Keywords:** Association Rules · Data Mining · Beam Pump Wells · Recommendation of Measures

## 1 Introduction

As the primary production equipment for domestic oil wells, beam pumps play a crucial role in petroleum production. However, issues such as rod parting, pump leakage, rod deviation and wear, and wax deposition significantly reduce the exempt period of beam pumps [1]. Therefore, it is of great significance to explore the intrinsic factors and patterns that affect the exempt period of oil wells, and provide corresponding recommended measures. This can effectively reduce the workload and extend the exempt period, contributing to improved operational efficiency.

The underground structure of oil wells is complex and constantly changing, with strong coupling of production parameters. The maintenance and management of beam pump wells lag behind, and there is an urgent need for production experience-supported information decision-making. Traditional methods for extending the exempt period are often based on empirical knowledge, without extensive utilization of historical real-time data feedback from well production. As a result, it is difficult to track the dynamic information of well production and identify the mixed effects caused by multiple factors. The perspective is limited to single-type problems. For example, in addressing the issue of rod wear, a directional lifting system was designed for a specific well, and no-rod lifting technology was adopted, which fundamentally solved the problem of rod wear [2]. However, this lifting system technology was designed based on the characteristics of a single block, and it has limitations and is difficult to be widely applied.

The advent of the era of big data in the petroleum industry has provided a vast stage for the application of data mining techniques. Accumulating massive historical data during long-term oil well development allows for the use of data mining techniques to process and analyze this data, uncovering valuable insights and experiences embedded within. Association rule mining, as one of the representative techniques in data mining, has made significant advancements in various areas such as disease recognition, drug prediction, and risk prediction of unsafe behaviors [3–5]. In this context, leveraging the mining capabilities of association rule algorithms for intrinsic factors of features promotes the deep integration of petroleum big data mining technology. To achieve this, a long-life well sample library was created based on historical production data from oil wells, combining expert experience with data processing techniques. By applying association rule mining algorithms, a frequent pattern library for long-life wells was constructed, and visual analysis was conducted on these frequent patterns.

The organizational structure of this paper is presented as follows: Sect. 2 introduces the methods for processing and selecting data features, as well as the techniques for association rule mining. Section 3 focuses on the frequent pattern mining of long-life wells in a specific oilfield, analyzing and discussing the frequent patterns and associated rules within these patterns. Section 4 presents the conclusions of the study and provides prospects for future research endeavors.

# 2 Preparation of the Sample Set for Beam Pump Well Exemption Period

## 2.1 Construction of Relevant Indicators System for Well Design

In response to the requirements of well design tasks, a comprehensive indicator system tailored for well design was developed by integrating expert knowledge. This indicator system consists of six major categories and includes over 100 parameters. The data sources for each parameter were identified, and a corresponding database was designed. The details of this indicator system are presented in Table 1.

**Table 1.** Presents the corresponding database for the well design study.

| Basic Information | Geological Data | Fluid Data | Mechanical and Production Data | Production Data | Operation Data |
|---|---|---|---|---|---|
| Well Number<br>Production Date<br>Oil (Gas) Field<br>Block Unit<br>Unit Name<br>Well Type<br>Lifting Method<br>… | Reservoir Type<br>Exploitation Layer<br>Effective Thickness<br>Saturation Pressure<br>Formation Temperature<br>Layer Porosity<br>Layer Permeability<br>Layer Saturation<br>… | Crude Oil Viscosity<br>Volume Coefficient<br>Water Mineralization<br>Formation Water Type<br>Freezing point<br>Wax Appearance Temperature<br>Wax Content<br>Gas-Oil Ratio<br>… | Structure Data<br>Tubing Data<br>Sucker Rod Data<br>Pump Data<br>Supporting Data<br>Pump Jack Data<br>Wellhead Data<br>Production Parameters<br>… | Daily Water Cut<br>Dynamic Liquid Level<br>Submergence Depth<br>Pump Efficiency<br>Power Consumption<br>System Efficiency<br>Indicator Diagram<br>Oil Pressure<br>… | Pump Testing Period<br>Exemption Period<br>Operation Time<br>Repair Causes<br>Construction Type<br>Failure Point Description<br>… |

Combining expert experience, a total of 15 features were selected for the oil well, including reservoir type, sand production, scale deposition, wax deposition, daily fluid production, normal water cut, dynamic liquid level, pump depth, submergence depth, stroke times, pump position wellbore inclination angle, salinity, crude oil viscosity, freezing point, and pump Size.

## 2.2 Data Integration and Standardization

Based on the operation big data of the pumping unit wells, standardized processing was carried out to address issues such as multiple data sources, varying frequencies, and mixed data types. This included the fusion of multi-source data, integration of data with different frequencies, and digitization of text-based data indicators. As a result, a standardized operation big data set for pumping unit wells was prepared. Refer to Table 2 for details.

**Table 2.** Operational sample set of beam pumping wells.

| Well | Number Reservoir | Category Sand | Depth Pump | Diameter Pump | Efficiency Water | Stroke | Stroke Count |
|---|---|---|---|---|---|---|---|
| 1 | Fault Block | No Sanding | 2000 | 44 | 53 | 3.05 | 2.4 |
| 2 | Medium to High Permeability | Slight Sanding | 1999 | 44 | 95 | 3.76 | 3 |
| 3 | Medium to High Permeability | Slight Sanding | 2003 | 44 | 96 | 3.71 | 2.2 |
| 4 | Complex Fault Block | No Sanding | 2100 | 44 | 95 | 3.88 | 2 |
| 5 | Complex Fault Block | No Sanding | 2000 | 44 | 68 | 3.41 | 2 |
| 6 | Complex Fault Block | No Sanding | 2002 | 44 | 21 | 3.85 | 2.7 |
| 7 | Complex Fault Block | Severe Sanding | 2007 | 38 | 98 | 2.96 | 2.5 |

## 2.3 Correlation Analysis

When two variables change in some degree as a result of each other's variations, we say they have correlation. Therefore, before data mining, analyzing the correlation between features and removing weakly correlated features can not only reduce workload but also improve model accuracy.

Common methods for correlation analysis include Pearson correlation coefficient, Spearman correlation coefficient, and Kendall correlation coefficient [6–8].

Pearson correlation coefficient formula:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E(X_{\ i} - \mu_X)E(Y_{\ i} - \mu_Y)}{\sigma_X \sigma_Y} \tag{1}$$

where $X_i$ and $Y_i$ represent the values of the $i$ observation, $\mu_X$ and $\mu_Y$ are the means of variables $X$ and $Y$ respectively, and $\sigma_X$ and $\sigma_Y$ are the standard deviations of variables $X$ and $Y$ respectively.

Spearman correlation coefficient formula:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{2}$$

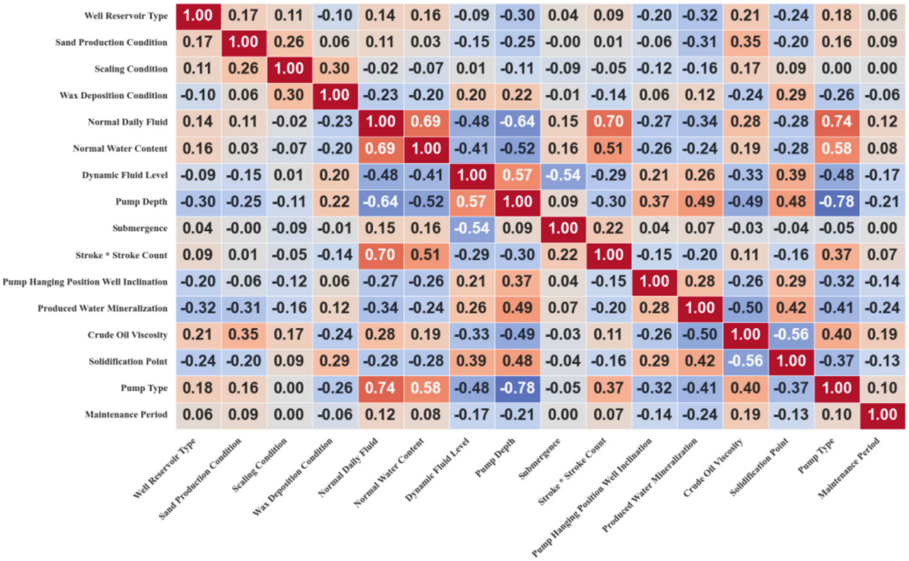
where $d_i$ represents the rank differences between the $i$ variable $X_i$ and $Y_i$, i.e., $X_i$ - $Y_i$, and $n$ represents the sample size.

Kendall correlation coefficient formula:

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} sgn\left(x_i - x_j\right) sgn\left(y_i - y_j\right) \tag{3}$$

where $sgn(x_i\text{-}x_j)$ and $sgn(y_i\text{-}y_j)$ represent the signs of rank differences between the $i$ and $j$ observations for variables $X$ and $Y$ respectively, and $n$ represents the sample size.

The Pearson correlation coefficient is commonly used for linear correlation analysis, the Kendall correlation coefficient is often used for comparing ordinal correlations, while the Spearman correlation coefficient can reflect both linear and nonlinear relationships between features. Therefore, Spearman correlation coefficient was chosen to analyze the relationship between features and the maintenance period, and the resulting correlation analysis is shown in Fig. 1.



| | Well Reservoir Type | Sand Production Condition | Scaling Condition | Wax Deposition Condition | Normal Daily Fluid | Normal Water Content | Dynamic Fluid Level | Pump Depth | Submergence | Stroke * Stroke Count | Pump Hanging Position Well Inclination | Produced Water Mineralization | Crude Oil Viscosity | Solidification Point | Pump Type | Maintenance Period |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Well Reservoir Type | 1.00 | 0.17 | 0.11 | -0.10 | 0.14 | 0.16 | -0.09 | -0.30 | 0.04 | 0.09 | -0.20 | -0.32 | 0.21 | -0.24 | 0.18 | 0.06 |
| Sand Production Condition | 0.17 | 1.00 | 0.26 | 0.06 | 0.11 | 0.03 | -0.15 | -0.25 | -0.00 | 0.01 | -0.06 | -0.31 | 0.35 | -0.20 | 0.16 | 0.09 |
| Scaling Condition | 0.11 | 0.26 | 1.00 | 0.30 | -0.02 | -0.07 | 0.01 | -0.11 | -0.09 | -0.05 | -0.12 | -0.16 | 0.17 | 0.09 | 0.00 | 0.00 |
| Wax Deposition Condition | -0.10 | 0.06 | 0.30 | 1.00 | -0.23 | -0.20 | 0.20 | 0.22 | -0.01 | -0.14 | 0.06 | 0.12 | -0.24 | 0.29 | -0.26 | -0.06 |
| Normal Daily Fluid | 0.14 | 0.11 | -0.02 | -0.23 | 1.00 | 0.69 | -0.48 | -0.64 | 0.15 | 0.70 | -0.27 | -0.34 | 0.28 | -0.28 | 0.74 | 0.12 |
| Normal Water Content | 0.16 | 0.03 | -0.07 | -0.20 | 0.69 | 1.00 | -0.41 | -0.52 | 0.16 | 0.51 | -0.26 | -0.24 | 0.19 | -0.28 | 0.58 | 0.08 |
| Dynamic Fluid Level | -0.09 | -0.15 | 0.01 | 0.20 | -0.48 | -0.41 | 1.00 | 0.57 | -0.54 | -0.29 | 0.21 | 0.26 | -0.33 | 0.39 | -0.48 | -0.17 |
| Pump Depth | -0.30 | -0.25 | -0.11 | 0.22 | -0.64 | -0.52 | 0.57 | 1.00 | 0.09 | -0.30 | 0.37 | 0.49 | -0.49 | 0.48 | -0.78 | -0.21 |
| Submergence | 0.04 | -0.00 | -0.09 | -0.01 | 0.15 | 0.16 | -0.54 | 0.09 | 1.00 | 0.22 | 0.04 | 0.07 | -0.03 | -0.04 | -0.05 | 0.00 |
| Stroke * Stroke Count | 0.09 | 0.01 | -0.05 | -0.14 | 0.70 | 0.51 | -0.29 | -0.30 | 0.22 | 1.00 | -0.15 | -0.20 | 0.11 | -0.16 | 0.37 | 0.07 |
| Pump Hanging Position Well Inclination | -0.20 | -0.06 | -0.12 | 0.06 | -0.27 | -0.26 | 0.21 | 0.37 | 0.04 | -0.15 | 1.00 | 0.28 | -0.26 | 0.29 | -0.32 | -0.14 |
| Produced Water Mineralization | -0.32 | -0.31 | -0.16 | 0.12 | -0.34 | -0.24 | 0.26 | 0.49 | 0.07 | -0.20 | 0.28 | 1.00 | -0.50 | 0.42 | -0.41 | -0.24 |
| Crude Oil Viscosity | 0.21 | 0.35 | 0.17 | -0.24 | 0.28 | 0.19 | -0.33 | -0.49 | -0.03 | 0.11 | -0.26 | -0.50 | 1.00 | -0.56 | 0.40 | 0.19 |
| Solidification Point | -0.24 | -0.20 | 0.09 | 0.29 | -0.28 | -0.28 | 0.39 | 0.48 | -0.04 | -0.16 | 0.29 | 0.42 | -0.56 | 1.00 | -0.37 | -0.13 |
| Pump Type | 0.18 | 0.16 | 0.00 | -0.26 | 0.74 | 0.58 | -0.48 | -0.78 | -0.05 | 0.37 | -0.32 | -0.41 | 0.40 | -0.37 | 1.00 | 0.10 |
| Maintenance Period | 0.06 | 0.09 | 0.00 | -0.06 | 0.12 | 0.08 | -0.17 | -0.21 | 0.00 | 0.07 | -0.14 | -0.24 | 0.19 | -0.13 | 0.10 | 1.00 |

**Fig. 1.** Spearman correlation analysis graph.

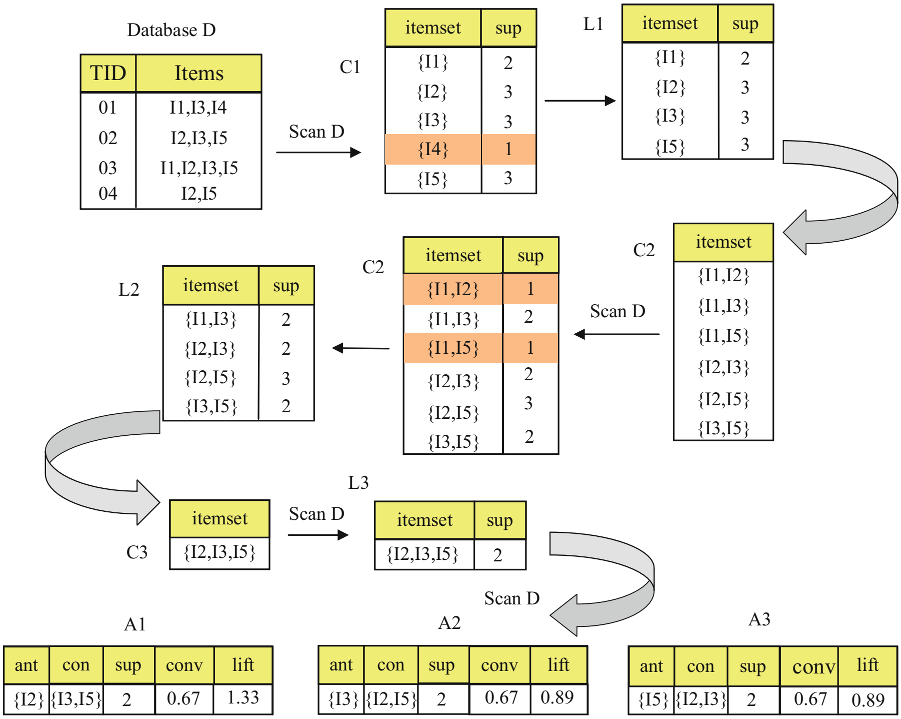
According to the correlation analysis graph in Fig. 1, it can be observed that the features affecting the maintenance period are, in sequence, Salinity, pump depth, and crude oil viscosity. Submergence and scaling condition have insignificant impact on the maintenance period.

# 3 Measures Recommendation Design for Long-Life Wells Based on Association Rules

## 3.1 Association Rule Mining Algorithm

Data mining algorithms can be used to discover frequent item-sets and association rules. In order to perform effective clustering, the A-priori algorithm [9–11] is employed for association rule mining. The strength of each rule is evaluated based on indicators such as support, confidence, and lift.

The A-priori algorithm generates frequent item-sets through the process of joining and pruning. The generated frequent item-sets are then used to generate association rules. As shown in Fig. 1, which illustrates the process of association rule mining, the algorithm starts by generating candidate item-sets (C1) based on the item categories in dataset D. Items below a certain threshold are removed, resulting in frequent 1-itemsets (L1). L1 is then combined to form 2-itemsets, generating candidate 2-itemsets (C2). Similarly, items below the threshold are removed to obtain frequent 2-itemsets (L2). This process continues, with L2 being combined to form 3-itemsets and generate candidate 3-itemsets (C3). Again, items below the threshold are removed to obtain frequent 3-itemsets (L3). Finally, the items contained in L3 are permuted and combined to form antecedents and consequents. Support, confidence, and lift are calculated for each rule (A1, A2, A3) based on the relationships between the antecedents and consequents (Fig. 2).



Database D

| TID | Items |
|---|---|
| 01 | I1,I3,I4 |
| 02 | I2,I3,I5 |
| 03 | I1,I2,I3,I5 |
| 04 | I2,I5 |

C1

| itemset | sup |
|---|---|
| {I1} | 2 |
| {I2} | 3 |
| {I3} | 3 |
| {I4} | 1 |
| {I5} | 3 |

L1

| itemset | sup |
|---|---|
| {I1} | 2 |
| {I2} | 3 |
| {I3} | 3 |
| {I5} | 3 |

C2

| itemset |
|---|
| {I1,I2} |
| {I1,I3} |
| {I1,I5} |
| {I2,I3} |
| {I2,I5} |
| {I3,I5} |

C2

| itemset | sup |
|---|---|
| {I1,I2} | 1 |
| {I1,I3} | 2 |
| {I1,I5} | 1 |
| {I2,I3} | 2 |
| {I2,I5} | 3 |
| {I3,I5} | 2 |

L2

| itemset | sup |
|---|---|
| {I1,I3} | 2 |
| {I2,I3} | 2 |
| {I2,I5} | 3 |
| {I3,I5} | 2 |

C3

| itemset |
|---|
| {I2,I3,I5} |

L3

| itemset | sup |
|---|---|
| {I2,I3,I5} | 2 |

A1

| ant | con | sup | conv | lift |
|---|---|---|---|---|
| {I2} | {I3,I5} | 2 | 0.67 | 1.33 |

A2

| ant | con | sup | conv | lift |
|---|---|---|---|---|
| {I3} | {I2,I5} | 2 | 0.67 | 0.89 |

A3

| ant | con | sup | conv | lift |
|---|---|---|---|---|
| {I5} | {I2,I3} | 2 | 0.67 | 0.89 |

**Fig. 2.** Illustrates the concept of association rule mining.

## 3.2 Preparation of Long-Life Well Measures Knowledge Base

The process measures supporting long-life oil wells in each block were statistically analyzed, and an association rule mining algorithm was used to create a process library for long-life wells. By setting the minimum support threshold to 0.01, a total of 104 frequent patterns for long-life wells were obtained, as shown in Table 3.

**Table 3.** Library of frequent patterns for long-life wells.

| rules | frequent item-sets |
|---|---|
| 1 | · Stroke * Stroke per Minute [>15]<br>· Wax Deposition Condition [No wax]<br>· Well Reservoir Type [Medium to high permeability]<br>· Water cut [>95]<br>· Pump Size [56 57]<br>· Dynamic Fluid Level [634–889]<br>· Salinity [0–10000]<br>· Angle of Inclination [0–15]<br>· Scaling Condition [Slight scaling]<br>· Crude Oil Viscosity [1000–10000]<br>· Sand Production Condition [Slight sand production]<br>· Submergence [100–300]<br>· Daily Fluid Production [30–80]<br>· Freezing point [null]<br>· Pump Depth [800–1100] |
| 2 | · Wax Deposition Condition [No wax]<br>· Stroke * Stroke per Minute [6.5–9.5]<br>· Well Reservoir Type [Medium to high permeability]<br>· Water cut [>95]<br>· Pump Size [56 57]<br>· Daily Fluid Production [10–30]<br>· Dynamic Fluid Level [634–889]<br>· Salinity [0–10000]<br>· Angle of Inclination [0–15]<br>· Scaling Condition [Slight scaling]<br>· Crude Oil Viscosity [1000–10000]<br>· Sand Production Condition [Slight sand production]<br>· Submergence [ 100–300]<br>· Freezing point [null]<br>· Pump Depth [800–1100] |

(*continued*)

**Table 3.** (*continued*)

| rules | frequent item-sets |
|---|---|
| 3 | · Wax Deposition Condition [No wax]<br>· Submergence [300–500]<br>· Stroke * Stroke per Minute [6.5–9.5]<br>· Well Reservoir Type [Medium to high permeability]<br>· Water cut [95-]<br>· Pump Size [56 57]<br>· Daily Fluid Production [10–30]<br>· Salinity [0–10000]<br>· Angle of Inclination [0–15]<br>· Scaling Condition [Slight scaling]<br>· Crude Oil Viscosity [1000–10000]<br>· Sand Production Condition [Slight sand production]<br>· Dynamic Fluid Level [306–634]<br>· Freezing point [null]<br>· Pump Depth [800–1100] |
| … | … |

Taking the first frequent pattern as an example, when the frequency of Stroke * Stroke per Minute is greater than 15, Wax Deposition Condition is “no wax deposition”, Well Reservoir Type is “medium to high permeability”, Water cut is greater than 95%, Pump Size is either 56 or 57, Dynamic Fluid Level is between 634–889 m, Salinity is within the range of 0–10000, Angle of Inclination is between 0–15 degrees, Scaling Condition is “slight scaling”, Crude Oil Viscosity is within the range of 1000–10000, Sand Production Condition is “slight sand production”, Submergence is between 100–300 m, Daily Fluid Production is within the range of 30–80, Pump Depth is between 800–1100 m, in this pattern, the wells exhibit long lifespan phenomena.

## 4 Measures Recommendation for Extending the Free Repair Period

Based on the historical data of the wells and expert consultation information, key indicators are determined. Then, the generated solutions from the historical data of the wells and scheduling rules are compared against the determined key indicators to identify expert experiences corresponding to similar well characteristics. Finally, a recommended action plan is formulated in accordance with the expert experiences.

For the target well, long-lived wells in the corresponding block are identified, and measures from these long-lived wells are recommended for the target well, as shown in Table 4.

**Table 4.** Example of Recommended Measures for Long-lived Wells.

| | Target Well | Recommended Wells |
|---|---|---|
| Well | Well 1 | Well 2 |
| Maintenance Period | 1 year | 5 years |
| Well Reservoir Type | Medium-high permeability | Medium-high permeability |
| Sand Production Condition | Slight sand production | Slight sand production |
| Scaling Condition | Slight scaling | Slight scaling |
| Wax Deposition Condition | Slight wax deposition | Slight wax deposition |
| Water cut | 30–90 | 30–90 |
| Freezing point | 0–30 | 0–30 |
| Pump Size | 56/57 | 56/57 |
| Crude Oil Viscosity | 1000–10000 | 1000–10000 |
| Daily Fluid Production (DFP) | 0–10 | 10–30 |
| Dynamic Fluid Level | 889–1260 | 0–306 |
| Pump Depth | 1100–1500 | 800–1100 |
| Submergence | 100–300 | >500 |
| Stroke * Stroke per Minute (SSPM) | 0–6.5 | 9.5–12.0 |
| Angle of Inclination | 15–30 | 0–15 |
| Salinity | 10000–50000 | 0–10000 |

Based on the geological features and process parameters between the target well and the candidate recommended well, the changes in lift values are analyzed based on shared characteristics. This process facilitates the identification of the optimal solution for process selection. As shown in Fig. 3, the lift values are compared for different processes between the target well and the recommended well. The lift values indicate the contribution to the long-term operational performance of the oil wells.

As shown in Fig. 3, when the geological indicators of oil wells, such as reservoir type, sand production, scale deposition, and wax deposition, are consistent, there are significant differences between the pump depth and output water salinity indicators of the target wells and the recommended wells. Therefore, we focus on analyzing the pump depth, submergence, and output water salinity indicators. The pump depth of the target wells ranges from 1100 m to 1500 m, the submergence ranges from 100 m to 300 m, and the output water salinity ranges from 10000 to 50000. In contrast, the recommended wells have a pump depth ranging from 800 m to 1100 m, submergence greater than 500 m, and output water salinity ranging from 1000 to 10000. As the output water salinity indicator is determined by geological conditions and is not easily changed, we can appropriately adjust the pump depth and submergence indicators to prolong the maintenance-free period of oil wells.

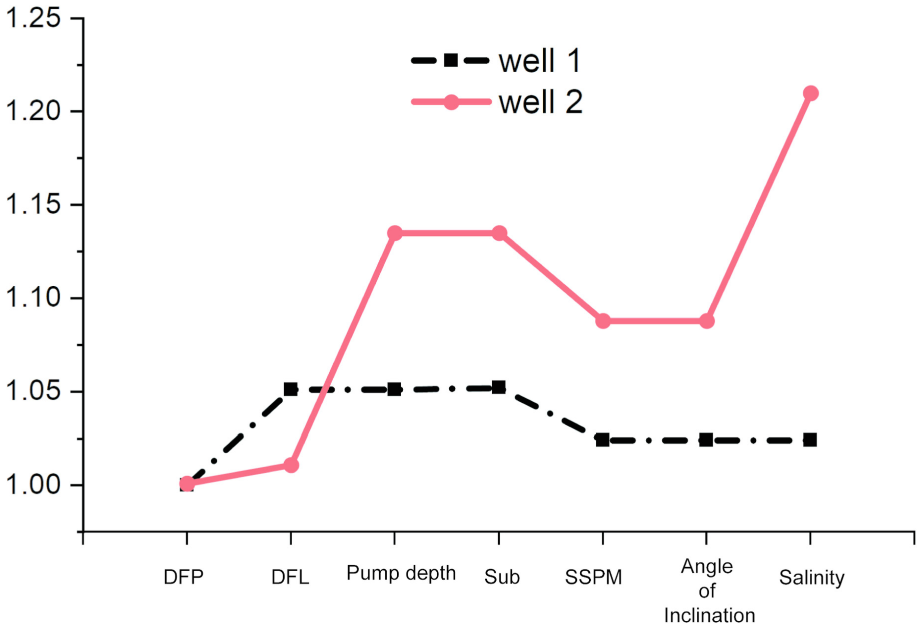


**Fig. 3.** Comparison of Oil Recovery Measures Indicators

## 5 Conclusion

Based on the comprehensive achievements of information construction in a certain oilfield, historical production data of 5,789 fully equipped medium-high permeability reservoirs with beam pumping wells were collected. Feature analysis samples and long-lived well samples were designed. Parameters were analyzed from multiple perspectives, including geology, fluid, production, lifting systems, and supporting processes. The focus was on comparing the differences between abnormal wells with extended maintenance intervals and regular wells in various parameters. Based on this analysis, factors and patterns influencing the extended maintenance interval were statistically summarized and identified.

After identifying the factors influencing the extended maintenance interval, the supporting patterns for long-lived wells were explored, resulting in the preparation of 105 frequent patterns for supporting long-lived wells. These patterns can be used to recommend measures and experiences for short-lived wells with similar geological characteristics, providing valuable guidance for the high-value application of oilfield big data.

In the future, efforts will be made to further expand the application of frequent patterns by integrating the recommended supporting patterns with field implementation codes, thereby reducing the difficulty of frontline application.

**Acknowledgments.** The project is supported by the Sinopec Scientific and Technological Research Project "Research on the Application of Big Data Technology in Oilfield Development" (No. P20071).

## References

1. Qun, L., Dingwei, W., Jianhui, L., et al.: Achievements and future work of oil and gas production engineering of CNPC. Pet. Explor. Dev. **46**(1), 145–152 (2019)
2. Qian, K., et al.: Research on fault diagnosis method of electric submersible plunger pump lifting technology. In: Lin, J. (ed.) Proceedings of the International Field Exploration and Development Conference 2021, pp. 5524–5537. Springer Nature, Singapore (2022). https://doi.org/10.1007/978-981-19-2149-0_506
3. Borah, A., Nath, B.: Identifying risk factors for adverse diseases using dynamic rare association rule mining. Expert Syst. Appl. **113**, 233–263 (2018)
4. Vougas, K., Sakellaropoulos, T., Kotsinas, A., et al.: Machine learning and data mining frameworks for predicting drug response in cancer: an overview and a novel in silico screening process based on association rule mining. Pharmacol. Ther. **203**, 107395 (2019)
5. Ruihua, Xu., Luo, F.: Risk prediction and early warning for air traffic controllers' unsafe acts using association rule mining and random forest. Safety Sci. **135**, 105125 (2021)
6. Porter, T.M.: Karl Pearson. Princeton University Press, Karl Pearson (2010)
7. Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson correlation coefficient. In: Cohen, I., Huang, Y., Chen, J., Benesty, J. (eds.) Noise reduction in speech processing, pp. 1–4. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-00296-0_5
8. Myers, L., Sirois, M.J.: Spearman correlation coefficients, differences between. Encycl. Stat. Sci. **12** (2004)
9. Chee, C.-H., Jaafar, J., Aziz, I.A., et al.: Algorithms for frequent itemset mining: a literature review. Artif. Intell. Rev. **52**(4), 2603–2621 (2018)
10. Ping-Hsun, Lu., Keng, J.-L., Kuo, K.-L., Wang, Y.-F., Tai, Y.-C., Kuo, C.-Y.: An apriori algorithm-based association rule analysis to identify herb combinations for treating uremic pruritus using Chinese herbal bath therapy. Evidence-Based Complement. Altern. Med. **2020**, 1–9 (2020)
11. Antomarioni, S., Ciarapica, F.E., Bevilacqua, M.: Association rules and social network analysis for supporting failure mode effects and criticality analysis: framework development and insights from an onshore platform. Safety Sci. **150**, 105711 (2022)

# Intelligent Diagnosis System for Oil Well Underground Conditions Based on Convolutional Neural Network

Jia-he Huang, Hong-hui Fan(✉), Wen-jie Liao, and Hui-ting Li

Jiangsu University of Technology, Changzhou, China
17314994669@163.com

**Abstract.** The existing pumping unit downhole working condition diagnosis system has a high false alarm rate and a low accuracy rate of diagnosis for complex working conditions and abnormal working conditions. To address this problem, a diagnostic system of pumping unit workings is developed. First of all, the displacement-load data of the workover diagrams were converted into images, and then, through preliminary screening, manual review and data balancing from hundreds of millions of workover diagrams accumulated over the years, a sample database of 28 types of workover conditions, such as normal production, insufficient fluid supply, gas influence, rod breakage and tubing leakage, was established, with a total of about 760,000 samples, to compile a data set that is leading in quality and quantity in China. The project adopts "graphic + data" composite diagnosis, "graphic" corresponds to the power diagram, "data" refers to the electrical parameters, set pressure and other production parameters, and transforms the fault diagnosis problem of the power diagram into a deep learning-based image classification problem. Deep learning based image classification problem. A fault diagnosis method based on migration learning and category imbalance loss is designed. Better diagnostic results are obtained, with the single diagnostic accuracy of no less than 98% for common working conditions, 99.5% for normal production, 98.4% for insufficient fluid supply, and 97.2% for gas influence.

**Keywords:** Convolutional neural network · fault diagnosis · oil extraction machine · transfer learning

## 1 Introduction

Currently, PCS, EPBP, and other information platforms in oil fields have brought great convenience to production management. However, the data collected through these platforms is not fully exploited. The load and displacement parameters generated during the reciprocating motion of the drilling head of a well machine are the parameters of dynamometer card. It reflects the impact of internal factors such as gas, oil, water, sand, and asphalt on the working condition of the drilling machine [1–3]. In the process of identifying the working condition of an oil well, dynamometer card fault diagnosis is an important method. Traditional dynamometer card fault detection is based on expert

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 205–214, 2024.
https://doi.org/10.1007/978-981-97-0272-5_17

systems and summarizes and analyzes different shapes of work diagrams. However, the actual dynamometer card are diverse and the relationship between formation reasons and faults is complex [2, 4–8]. Overreliance on expert knowledge leads to costly and time-consuming system development and low robustness. The accuracy of graph description is low and some characteristic features of the work diagram are difficult to describe. Existing downhole working condition diagnosis systems have high warning misreporting rates, low accuracy in diagnosing complex and abnormal working conditions, and lack of targeted solution after diagnosis, requiring technical personnel to develop solutions again.

In recent years, the development of machine learning technology has made it possible to achieve high-precision recognition of oil well conditions using massive data [9, 10]. Artificial neural networks, BP neural networks, and self-encoders are among the machine learning methods applied in oil well performance diagnostics, and all have achieved good results. Based on SVM and other classification models, good results can also be achieved, but the condition is that reasonable feature selection is conducted. Deep learning has gradually emerged as a popular method, and the most representative image recognition method is convolutional neural networks. Similar to deep learning, it also requires a certain level of domain knowledge and experimental analysis [10–12]. Since the arrival of the GPU computing era, many professionals have attempted to explore the application of convolutional neural networks in performance diagnostics for oil wells.

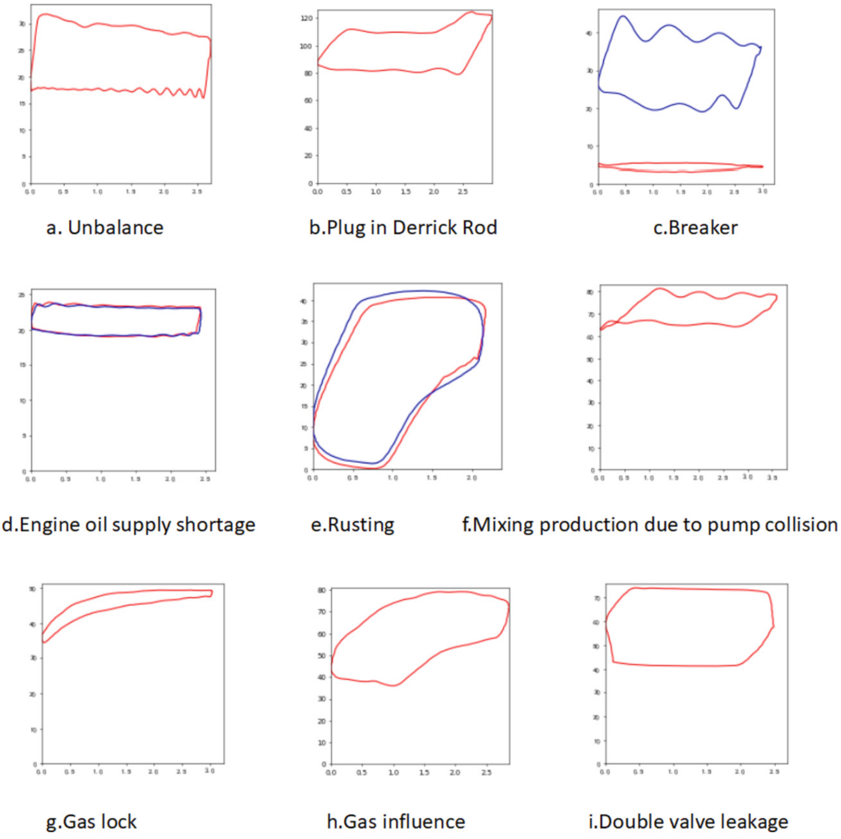
Research priorities for using CNN to diagnose oil extraction machinery underground conditions include: It is necessary to focus on studying the problems of high warning false positive rates and low diagnostic accuracy for complex and abnormal operating conditions of underground oil extraction conditions in existing oil extraction condition diagnosis systems, and effective extraction of the outline features of the performance graph, as well as strengthening the extraction of highly discriminating outline features. Therefore, a set of deep learning-based and expert experience fusion performance graph real-time diagnosis technology is developed to transform the fault diagnosis problem of the performance graph into an image classification problem based on deep learning. The loss function FocalLoss is used to alleviate the class imbalance problem and enhance the learning of the model for rare samples, solving the problem of poor diagnostic accuracy for rare data. The CAM method is used to present the parameters of the convolutional neural network in a visually intuitive hot region format in the final diagnostic image, so that the model can be judged whether it has sufficient attention to the key regions of the performance graph that cause incorrect classification. This can solve the problem of high warning false positive rates and low diagnostic accuracy in existing diagnosis systems and intelligently provide strategic solutions to guide on-site production.

## 2 Production Data Analysis

### 2.1 Dynamometer Card Analysis

The extraction rod of the oil extraction machinery moves from the bottom to the top, and then back to the bottom, repeating this process. This process is called the stroke. During one stroke of the extraction rod, the point of suspension's displacement and load will form a closed figure, which is called the dynamometer card.

Using actual workload graphs to diagnose oil well conditions is a widely adopted method in oil field production. The different geometric shapes of the workload graph represent different operating conditions of the well. Actual workload graphs can reflect abnormal operations of deep well pumps in underground conditions. Combining geological conditions, production data, and the condition of instruments to analyze and interpret the working system of the oil well and the compatibility of the machine, rod, and pump parameters with the well. Figure 1 are examples of dynamometer card.



a. Unbalance b.Plug in Derrick Rod c.Breaker

d.Engine oil supply shortage e.Rusting f.Mixing production due to pump collision

g.Gas lock h.Gas influence i.Double valve leakage

**Fig. 1.** Example of dynamometer card

## 2.2 Electrical Parameter Analysis

During the operation of the well, the three-phase electrical parameters that can be collected by the electrical control cabinet include RMS current, RMS voltage, active power and reactive power. The average active power and the coefficient of active power fluctuation can be used as parameters for well condition diagnosis, and the average active power is calculated as follows:

$$P_m = \frac{1}{T_0} \int P(t)dt \tag{1}$$

Calculate the coefficient of active power fluctuation:

$$\mathrm{K} = \frac{\sqrt{\frac{1}{T_0} \int P^2(t)dt}}{\frac{1}{T_0} \int P(t)dt} \tag{2}$$

$\mathrm{T}_0$ represents the electrical parameter collection cycle, with a set collection cycle of 2 h, and $(t)$ is the active power. During the normal working process of the pumping unit, the average load rate of each stroke of the pumping unit is stable and stable in a straight line. The load rate will be affected by the operating conditions of the oil well, but the average active power $\mathrm{P_m}$ fluctuates steadily within a reasonable range. The fluctuation coefficient K of active power can reflect the fluctuation of the pumping unit load within a stroke, and the larger K, the poorer the balance and stability of the pumping unit. The balance rate of the pumping unit can be determined based on the collected electrical parameters, thereby understanding the changes in downhole load of the oil well.

### 2.3 Data Preprocessing

In Jiangsu oil field, the exploration and production units of Unit 1 and Unit 2 of the drilling production plant have a total of nine management areas and nearly 2,000 wells. From 2015 to 2020, almost one hundred million pictures of dynamometer card have been accumulated. In order to obtain samples of work efficiency diagrams of normal and various fault conditions, it is necessary to select from this billion-plus number of pictures, and the following work was implemented:

(1) Preliminary screening: A preliminary screening program was written based on the principle of small batches and elimination of obvious duplicates, to screen work efficiency data. The traditional work efficiency diagram diagnosis method was used to iterate through all the historical dynamometer card of all management areas, and the work efficiency data of different diagnoses was drawn as an image and saved in different directories.
(2) Human review: A large amount of work efficiency data was labeled for operating conditions after screening, and human review was performed on each dynamometer card. Dynamometer card with incorrect diagnoses were assigned to the appropriate directories for faulty operations, and dynamometer card with unidentified faults were submitted for review by oil field experts.
(3) Data balance: The number of work efficiency diagrams corresponding to different fault types was very imbalanced. There were many work efficiency diagrams for normal and insufficient fluid operations, while the number of diagram types such as cylinder lock pump cylinder, fixed valve leakage, and mobile valve leakage was very small. This had a significant impact on the results of artificial intelligence data training, significantly reducing the accuracy of diagnosis. In response to this

situation, we wrote a data reduction program to delete numerous and similar work efficiency diagrams. For rare fault types with few diagram samples, a separately written enhancement program was written to screen from historical work efficiency diagrams again. This program had a higher tolerance rate than the previous screening program. The results were then reviewed by humans.

# 3 Algorithm Research and Model Design

## 3.1 Algorithm Research

Current methods for fault diagnosis using dynamometer card generally have the following drawbacks:

1. Convolutional neural network as a deep learning model have inherent black-box properties. Existing CNN-based work efficiency diagram diagnosis models can only reflect the diagnosis results of a specific work efficiency diagram in the final classification probability, without specific visualization of the many parameters within the network. This makes it difficult to evaluate the effectiveness of the model from a professional experience perspective.
2. The accuracy of recognition by deep learning models is dependent on the quality of the data set. The larger the number of categories in the data set, the better the training effect will be. In the context of work efficiency diagram diagnosis for oil wells, such as fixed leakage and mobile leakage, due to their rarity, there are few samples, resulting in a long tail distribution across the entire data set, which has a very negative impact on the training and final diagnosis accuracy of deep learning models. However, existing CNN-based work efficiency diagram diagnosis models do not make adjustments to address the problem of long-tail distribution in data sets, and therefore have poor performance in diagnosing rare data.
3. The performance of deep learning models also depends on the good initialization of network parameters. Some existing work efficiency diagram diagnosis models typically use a random initialization method to set the initial parameters of the network. This can affect the final convergence of the model to a certain extent, thereby reducing the performance of fault diagnosis.

Based on the above three issues, a method for fault diagnosis using transfer learning and class imbalance loss was designed. Transfer learning technology, as the mainstream initialization method in the field of deep learning, uses parameters trained on a large image data set ImageNet as the initial parameters of the network and performs parameter fine-tuning using work efficiency diagram data. This addresses the third issue. For the second issue, a loss function called Focal Loss was used to alleviate the problem of class imbalance, in order to improve the learning level of rare samples for the model. For the first issue, the CAM method was used to visualize the parameters of the convolutional neural network in a thermal map format in the final diagnosis image, allowing the model to focus on key regions of work efficiency diagrams that were incorrectly classified.

### 3.2 Model Design

The diagnosis of dynamometer card falls into the category of N-class image classification, where N is the number of categories. The normal, imbalance, gas influence, insufficient fluid, and valve failure categories are all included in the 26 categories of work efficiency diagrams. The routine work efficiency data is stored in a database in the form of binary coding, which is decoded by a Python program and drawn on a canvas to form an image. The axis of the horizontal and vertical coordinates is retained to provide scale information for model recognition, and the image is saved with a resolution of 224 x 224 on local storage.

The entire model's output is extracted through an intermediate pre-trained model for feature extraction. The pre-trained model contains multiple residual blocks within its internal structure. The feature vectors are then transformed into probability distributions for target classifications through a fully connected layer, achieving model prediction. Finally, the probabilities are normalized using the softmax activation function to obtain the final classification results.

This article uses the SeResnet50 network as the model framework for training, where the prefix "Se" stands for the squeeze and excitation process. This process involves adding a SE module to the ResNet50 network model. The SE (Squeeze-and-Excitation) module adaptively redefines each channel's feature by separately modeling information from each feature layer. The module is not a complete network structure but rather a sub-structure that can be nested into other classification or detection models. The principle of the process is to enhance important features and weaken unimportant ones by controlling the scale of the SE module, similar to the mechanism of attention. The process is mainly aimed at making the extracted features more pointing, thus better recognizing fine features in FGVC tasks (Fig. 2).

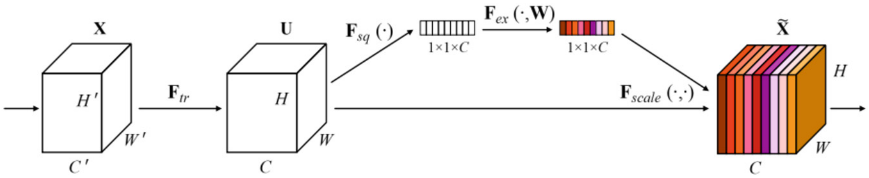


**Fig. 2.** SE module structure diagram

### 3.3 Accuracy Analysis

The CAM algorithm is used to visualize the features of samples with insufficient fluid supply. The characteristics of insufficient fluid supply are that the weight of the hanging point cannot decrease immediately during the downward stroke, and only when the piston contacts the fluid surface can it be rapidly unloaded. This is reflected in the load-deflection curve, as shown in Fig. 3, where a missing corner should be present in the lower right corner of the load-deflection curve. After visualizing the discrimination features learned by our model, it can be seen that our model focuses on the missing corner point and learns the discrimination features for load-deflection curve classification. Through the visualization of the CAM algorithm, it better reflects the detection hotspots and interpretability of the model for various types of load-deflection curve failures. In response to the problem of long tail distribution, the loss function we use, the Focal Loss function, produces a large loss value for samples with fewer categories, and a smaller loss value for samples with more categories, significantly improving the accuracy of classification for samples with few categories and reducing the negative effect brought about by long tail distribution, resulting in an increase in final model accuracy of 6%.

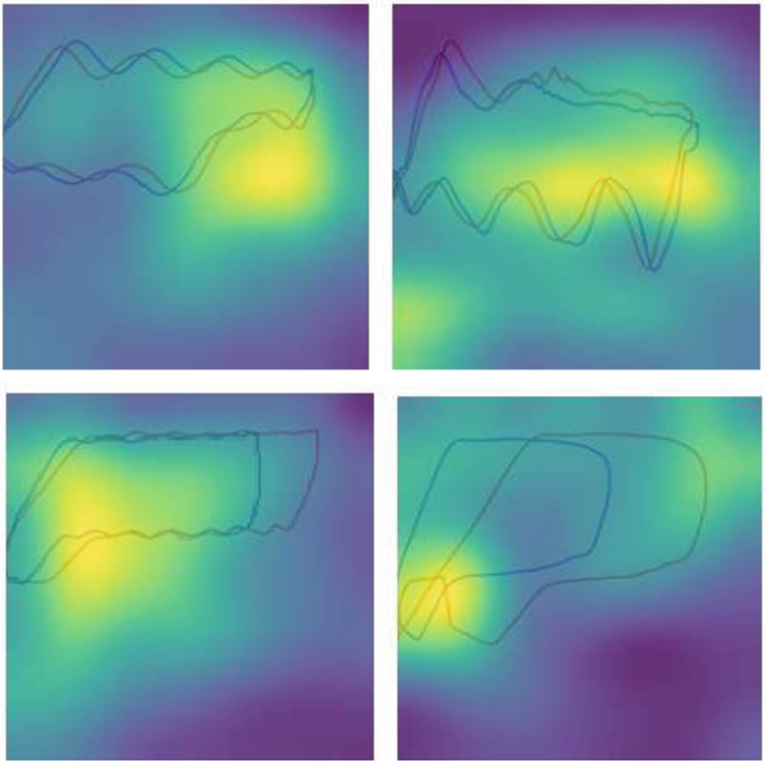


**Fig. 3.** Applying CAM methods to interpret dynamometer card

## 4 Experiments and Results Analysis

This article explores the effectiveness of different residual networks as feature extractors, including ResNet, DenseNet, and MobileNet. Among the three models, ResNet shows the strongest robustness to imbalanced data, while DenseNet performs the poorest. This is because the strong generalization ability brought about by the extensive connections in

DenseNet is actually achieved by ignoring the features of a few rare samples. MobileNet is more concise and therefore achieves better results, but its Transfer Learning is weaker than ResNet's, so it naturally trails behind in low-sample situations. Figure 4 shows a comparison of experiments on seven typical working conditions for the three models.

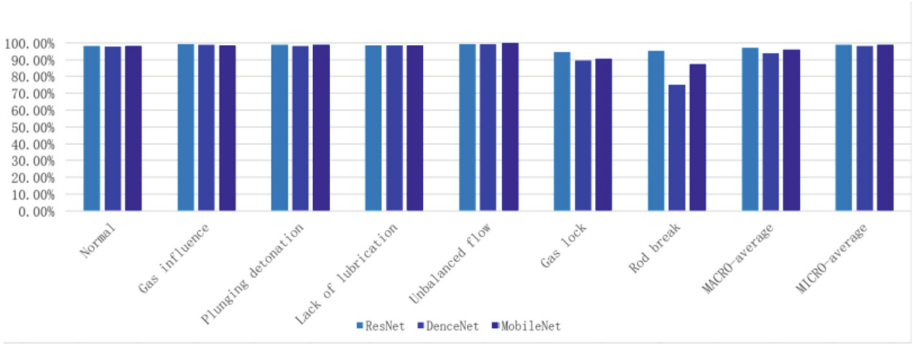


**Fig. 4.** Experimental comparison of seven typical working conditions on three models

Table 1 shows the experimental results of the resnet model.

**Table 1.** Experimental results of the resnet model.

| No | Type of working condition | Precision | Recall | F1-score | Support |
|---|---|---|---|---|---|
| 1 | Downside obstruction | 0.96 | 0.94 | 0.95 | 144 |
| 2 | Imbalance | 1.00 | 1.00 | 1.00 | 3853 |
| 3 | Insufficient fluid supply | 0.99 | 1.00 | 1.00 | 31522 |
| 4 | Power diagram error | 0.99 | 0.99 | 0.99 | 12381 |
| 5 | Fixed valve normally open | 1.00 | 0.97 | 0.98 | 59 |
| 6 | Fixed Valve Leakage | 0.99 | 0.99 | 0.99 | 964 |
| 7 | Pumping and spraying | 1.00 | 1.00 | 1.00 | 822 |
| 8 | Pumping rod on touch | 0.98 | 0.98 | 0.98 | 2166 |
| 9 | Rod broken | 0.96 | 0.98 | 0.97 | 46 |
| 10 | Plunger stuck | 1.00 | 1.00 | 1.00 | 3 |
| 11 | Plunger out of pump barrel | 0.93 | 0.94 | 0.93 | 113 |
| 12 | Normal | 1.00 | 1.00 | 1.00 | 41278 |
| 13 | Gas influence | 0.99 | 1.00 | 1.00 | 1065 |
| 14 | Airlock | 0.00 | 0.00 | 0.00 | 1 |

(*continued*)

**Table 1.** (*continued*)

| No | Type of working condition | Precision | Recall | F1-score | Support |
|---|---|---|---|---|---|
| 15 | Oil well out of sand | 1.00 | 0.92 | 0.96 | 26 |
| 16 | Oil pipe leak or well wash | 0.99 | 1.00 | 0.99 | 1441 |
| 17 | Swim valve normally open | 0.99 | 1.00 | 0.99 | 414 |
| 18 | Leaky pilot valve | 0.98 | 0.98 | 0.98 | 227 |
| 19 | Pump production | 0.98 | 0.98 | 0.98 | 3467 |
| 20 | Wax formation | 0.99 | 1.00 | 1.00 | 4278 |

### 4.1 Practical Applications

As of March 25, 2022, the types of working conditions diagnosed based on dynamometer card fault diagnosis were divided into 26 categories, and a total of 19 types of faults occurred in the nine management areas of Jiangsu Oilfield Plant I and Plant II during this period, select the three most common to display, which were manually reviewed and counted as follows (Table 2):

**Table 2.** Results of the three most common types of troubleshooting.

| NO | Type of working condition | Number of oil wells | Number of misreported wells | Number of wells missed | Correct rate |
|---|---|---|---|---|---|
| **1** | Normal | 798 | 2 | 2 | 99.5% |
| **2** | Insufficient fluid supply | 618 | 0 | 10 | 98.4% |
| **3** | Gas influence | 180 | 5 | 0 | 97.2% |

## 5 Conclusions

The well condition diagnosis based on the workup graph uses the pre-training-fine-tuning paradigm of migration learning, combined with the loss function FocalLoss for category imbalance, to reduce the negative effect of the long-tail distribution and significantly improve the performance of the model in the workup graph fault diagnosis of oil wells. Using the CAM method, the parameters in the convolutional neural network are presented in the final diagnostic map in the form of visualized hot zones, from which it can be judged whether the model has sufficient attention to the key areas of the schematic power map misclassification. Better diagnostic results were obtained, with a single diagnostic accuracy of no less than 95% for common operating conditions, 99.5%

for normal production, 98.4% for insufficient fluid supply, and 97.2% for gas influence. Improved response timeliness, with an average reduction of 4 h in fault response time. Improved management, timely handling of faults, avoiding complications of faults, timely maintenance, and improved equipment integrity.

## References

1. Hao, D., Gao, X., Li, X.: Motor Power Based Inversion of Dynamometer Cards Using Hybrid Model
2. Tian, Z., Li, K., Gao, X., et al.: Status Quo of Research on the Application of Dynamometer Card in Oil Production Process Control
3. Eisner, P., Langbauer, C., Fruhwirth, R.K.: Comparison of a novel finite element method for sucker rod pump downhole dynamometer card determination based on real world dynamometer cards. Upstream Oil Gas Technol. **9**, 100078 (2022)
4. Xiaoxiao, L., Wang, H., Liu, Y., Chen, S., Lan, W., Sun, B.: A novel method of output metering with dynamometer card for SRPS under fault conditions. J. Pet. Sci. Eng. **192**, 107098 (2020)
5. Boyuan, Z., Gao, X., Li, X.: Diagnosis of sucker rod pump based on generating dynamometer cards. J. Process Control **77**, 76–88 (2019)
6. Wei Jingliang, Gao Xianwen. Electric-parameter-based inversion of dynamometer card using hybrid modeling for beam pumping system. Math. Probl. Eng. **2018**, 6730905 (2018)
7. Tao Ren, Xiaoqing Kang, Wen Sun, Hong Song. study of dynamometer cards identification based on root-mean-square error algorithm. Int. J. Pattern Recogn. Artif. Intell. **32**(2), 1850004 (2018)
8. Yan, N., Dai, S.J.: Research on Displacement Calculation of Dynamometer Card Based on Kalman Filter and Discrete Numerical Integration
9. Jürgen, S.: Deep learning in neural networks: an overview. Neural Netw, **61**, 85–117 (2015)
10. Alex, K., Sutskever, I., Geoffrey, E.H.: ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. **25**, 1097–1105 (2012)
11. Szegedy, C., Liu, W., Jia, Y., et al.: Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). USA, pp. 14071. IEEE (2014)
12. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). USA. IEEE (2016)

# Intelligent Kick Detection Method Using Cascaded GRU Network with Adaptive Monitoring Parameters

De-zhi Zhang[1], Wei-feng Sun[2](✉), Yong-shou Dai[2], Sai-Sai Bu[2], and Jian-han Feng[2]

[1] College of Control Science and Engineering, China University of Petroleum (East China),
Qingdao, China

[2] College of Oceanography and Space Informatics, China University of Petroleum (East China),
Qingdao, China

sunwf@upc.edu.cn

**Abstract.** Drilling sensor failure leads to unavailability of kick monitoring parameters and the inability to apply intelligent kick detection methods. To solve this problem, a confidence evaluation indicator based on softmax is designed to measure the difficulty of kick identification, and the appropriate monitoring parameters are adaptively selected based on this indicator. Finally, an intelligent kick detection method using a cascaded GRU network with adaptive monitoring parameters is proposed in this paper. Kick identification experiments were conducted using simulated and measured data. The experimental results show that, when one monitoring parameter is unavailable, the recognition accuracy of the cascaded network proposed is improved by 10.61% on average and the computational load is reduced by 38.5% compared with the traditional gate recurrent unit network. The applicability of intelligent kick detection methods is significantly improved.

**Keywords:** Intelligent Kick Detection · Difficulty of Kick Identification · Adaptive Monitoring Parameters · Gate Recurrent Unit

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 215–224, 2024.
https://doi.org/10.1007/978-981-97-0272-5_18

## 1 Introduction

Kick is the phenomenon of formation fluids influx into the borehole, caused by formation pore pressure being higher than the bottom hole pressure (BHP). Kick can cause a blowout accident, resulting in huge casualties and economic losses. Therefore, accurate early kick detection has always received high attention from drilling experts.

Many studies use multi-parameters for kick detection. The monitoring parameters include mud parameters, engineering parameters, and downhole parameters. Mud parameters that reflect the change in total mud volume include differential flow out (DFO) and pit volume [1]. Using them to detect kicks will make some delay. Engineering parameters include rate of penetration (ROP), weight on bit (WOB), torque, pump pressure, standpipe pressure (SPP), hook load [2], etc. They have fast responses, but they are prone to receive interference from human operation and environmental factors, resulting in low accuracy of kick detection. Downhole parameters are collected from logging-while-drilling (LWD) and measurement-while-drilling (MWD), including downhole annular pressure, drilling fluid density, drilling fluid conductivity [3], etc. It is more accurate and timely kick detection using downhole parameters because their collection location is close to the location of kick occurrence. However, downhole acquisition equipment is still not widely used in some well sites due to its expensive price, slow transmission efficiency, and poor equipment stability.

The existing kick detection methods include model-based methods and data-based methods. Major oil service companies, including Schlumberger [4], Halliburton [5], Shell International Exploration and Production [6], and Corva [7], have mainly studied model-based kick detection methods. The model-based method needs to build a mathematical model of pit volume and drilling fluid flow, then alarm when two parameter values exceed the preset safety thresholds. The more reasonable the mathematical model constructed, the higher the accuracy of kick identification. However, there are many assumptions in the modeling process, as well as some model parameters and safety thresholds are difficult to set. So, the accuracy of the model-based method needs to be improved. In contrast, data-based monitoring methods have self-learning capability, thus reducing the number of missed and false alarms caused by unreasonable settings of model parameters and thresholds. Therefore data-driven kick detection models often use multiple parameters as input to achieve earlier kick detection. Many universities and researchers use data-driven methods to achieve earlier kick detection, including support vector machines (SVM) [8], Bayesian networks (BN) [9], random forests (RF) [10], and Artificial Neural Networks (ANN) [11]. In addition, deep learning techniques, such as Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) [12], show greater advantages in processing drilling time series data. Because it is more accurate to identify the kick by analyzing the trend of the parameters than by the parameter values.

The studies by major oil service companies prove that most kicks can be monitored based on anomalous changes in pit volume and flow-out rate, and the research by universities indicates that using engineering parameters can achieve earlier kick detection. However, engineering parameters are easy to become unavailable due to environmental factors in the field. And the data-driven kick detection methods are only useful if all input parameters are available. In order to accurately detect kicks using mud parameters and engineering parameters and to improve the robustness of the kick detection method, an intelligent kick detection method based on a cascaded gated recurrent unit (GRU) network with adaptive monitoring parameters is proposed. The cascaded GRU network consists of two sub-networks: a basic GRU network and a comprehensive GRU network. The basic network uses the mud parameters to identify most of the kicks. The comprehensive network uses the engineering parameters and the results of the basic network as input to achieve the monitoring of difficult-to-identify kicks. In order to automatically distinguish the kicks with different identification difficulties, a confidence evaluation indicator is designed using the softmax algorithm to evaluate the reliability of the output of the basic network. The results of the basic network can also be applied in the comprehensive network, thus avoiding the increase of computational load. Therefore, the method not only has fewer calculations but also is more robust.

## 2 Method

The proposed method aims to use only mud parameters (pit volume and DFO) to monitor most of the kicks and to automatically use both mud parameters and engineering parameters (ROP, WOB, torque, and SPP) to monitor a few difficult-to-identify kicks. The proposed intelligent kick detection method consists of three sequentially connected modules: a basic model, a switching module, and a comprehensive model (see Fig. 1).

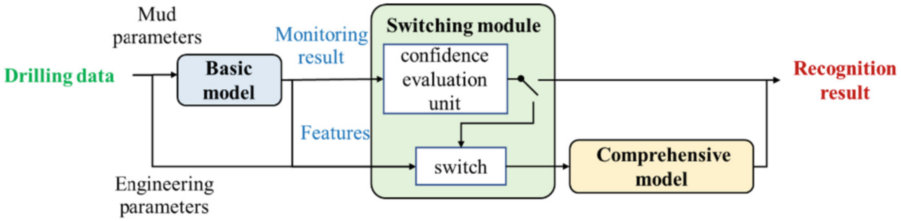


**Fig. 1.** Kick detection process of proposed method.

First, a kick is monitored by the basic model constructed with the mud parameters as input. Then a confidence evaluation unit designed in the switching module is used to determine which kicks can be accurately monitored by the basic model. Where the confidence evaluation indicator is calculated by the softmax algorithm [13] and is used as a "switch" to control the use of the comprehensive model.

$$p(y|X) = \frac{e^{g(X,y)}}{\sum e^{g(X,Y)}} \tag{1}$$

where $p$ is the indicator, $X$ denote the inputs, $y$ denotes the outputs for each type, $Y$ denote the outputs for all types, $e$ denotes an exponential function, $g$ denotes the basic model.

When the indicator value is lower than the preset threshold, the comprehensive model performs. The comprehensive model uses the engineering parameters and the features of mud parameters as inputs. This structure avoids repeated calculations. Therefore, the automatic selection of monitoring parameters in this method not only has less computational load but also avoids the reduction of accuracy when the engineering parameters are unavailable. Thus, the robustness of the intelligent kick detection method is improved. Both the basic and comprehensive models are based on the GRU, which can realize the feature extraction and classification of time series.

### 2.1 GRU

GRU is a deep learning technique based on the LSTM simplified architecture. It also could analyze the time series and avoid the long-term dependencies in RNN. The basic structure of GRU consists of two gating units (a reset gate and an update gate) (see Fig. 2). It is easier to train and compute than the three-gating unit structure of LSTM network (a forgetting gate, an input gate, an output gate) [14]. To prevent important historical information from being forgotten, the update gate in the GRU module filters and records historical information, while the reset gate analyzes the relationship between historical information and the current input. Where $M$ is the recurrent module of GRU, $L$ is the fully connected layer, which classifies the time series according to the state of the hidden layer at the last moment.

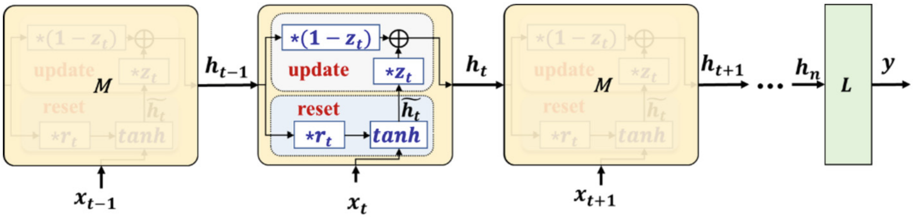


**Fig. 2.** Process of GRU network processing time series.

The formula for each parameter in the GRU network.

$$z_t = \sigma(w_z * [h_{t-1}, x_t] + b_z) \tag{2}$$

$$r_t = \sigma(w_r * [h_{t-1}, x_t] + b_r) \tag{3}$$

$$\tilde{h}_t = \tanh(w_h * [r_t * h_{t-1}, x_t] + b_h) \tag{4}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t) \tag{5}$$

$$y = \sigma(w_o * h_n + b_o) \tag{6}$$

where the values of weights $(w_z, w_r, w_h, w_o)$ and biases $(b_z, b_r, b_h, b_o)$ are obtained by training using a backpropagation algorithm.

### 2.2 Cascaded Network

The cascaded network consists of two sequentially connected GRU networks, a basic GRU network and a comprehensive GRU network (see Fig. 3). The basic GRU network uses DFO and pit volume (PitV) as inputs, while the comprehensive GRU network employs ROP, WOB, torque (Tor) and SPP as inputs. The trend features of the time series are obtained in the GRU module *M*, then the concatenate layer *C* combines all features, and the fully connected layer *L* classifies the time series based on these features. The basic GRU network classifies the drilling data according to the trend features of DFO and pit volume, while the comprehensive network depends on the trend features of DFO, pit volume, ROP, WOB, torque, and SPP.

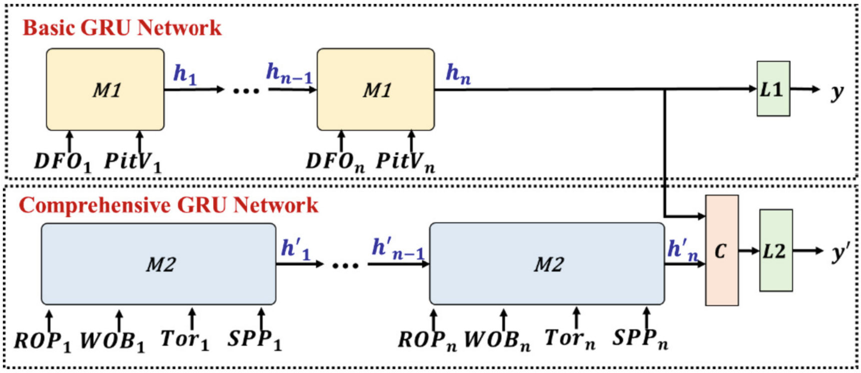


**Fig. 3.** Structure of a cascaded GRU network.

The following Table 1 gives the number of nodes in each layer of the cascaded GRU network.

**Table 1.** Number of nodes in each layer of the cascaded GRU network.

| Layer name | Basic GRU network | Comprehensive GRU network |
|---|---|---|
| Input layer | (60,2) | (60, 4) |
| GRU module | 32 | 64 |
| Concatenate layer | / | 96 |
| Fully connected layer | 8 | 24 |
| Output layer | 2 | 2 |

## 3 Data

The drilling data used in this paper were collected from 29 wells in Shengli oilfield, and a total of 35 kicks occurred. The monitoring parameters, including ROP, WOB, torque, SPP, DFO, and pit volume, are available in 20 wells. In addition, some parameters in

9 wells have missing data, and the reasons include sensor not working or equipment failure. The following Table 2 gives the number of kick cases with missing parameters and the type of missing parameter.

**Table 2.** Number of kick cases with missing parameters.

| Missing parameter | Number of kicks |
|---|---|
| None | 24 |
| ROP | 9 |
| WOB | 5 |
| Torque | 9 |
| SPP | 3 |
| DFO | 1 |
| Pit volume | 0 |

There are many outliers, missing values, and noise in the measured drilling data. So, before training the intelligent detection model with the measured kick data, data pre-processing, and sample set construction are required. Data pre-processing including outlier removal, missing value processing, noise removal, and normalization could improve the quality of training data.

The 3 sigma method [15] is used to identify outliers and then remove them.

$$x \in \{|x - \mu| > 3\sigma\} \tag{7}$$

where $\mu$ denotes the mean and $\sigma$ denotes the standard deviation.

A linear interpolation method [16] was used to fill in the individual missing values.

$$x_i = \frac{x_{i-t} + x_{i+t}}{2t} \tag{8}$$

where $t$ indicates the time interval between the missing data and the adjacent valid data.

Then, a time series denoising method based on the heat conduction equation [17] was used to remove noise from the data.

$$u_i^{j+1} = u_i^j + \lambda(u_{i-1}^j + u_{i+1}^j - 2u_i^j) \tag{9}$$

where $u$ denotes the signal values, $i$ represents the sampling instant, $j$ denotes the number of iterations, and $\lambda$ is the conduction rate.

Finally, the data are mapped to the interval [0,1] using min-max normalization [18].

$$\dot{x} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{10}$$

where $x_{max}$ denotes maximum value, and $x_{min}$ denotes minimum value.

A sample refers to a time series of monitoring parameters, including ROP, WOB, torque, SPP, DFO, and pit volume. These time series were obtained using the window slicing method. The length of the time window is 5 min. A total of 17,774 samples are obtained, of which 14,954 samples without missing parameters are divided into the training and testing sets in the ratio of 8:2. 2,820 samples with missing parameters are used as the validation set to check the robustness of the model. The sample labels are marked by experts based on drilling records.

# 4 Result and Discussion

The accuracy of the kick detection model is calculated using a confusion matrix. The following Table 3 gives a confusion matrix including True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The accuracy is calculated as

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{11}$$

**Table 3.** A confusion matrix for kick detection.

| Predicted | Actual | |
|---|---|---|
| | Kick | No Kick |
| Kick | TP | FP |
| No kick | FN | TN |

First, the value of the threshold in the confidence evaluation unit needs to be preset. Different values of the threshold correspond to different accuracies of the kick detection model (see Fig. 4). The accuracy of the testing set increases as the threshold increases, while the accuracy of the validation set decreases as the threshold increases. So, the accuracy of testing and validation sets increases and then decreases. When the threshold value is 0.8, the accuracy of the test set and validation set is maximum.

Traditional GRU network is built with DFO, pit volume, ROP, WOB, torque, and SPP as inputs. For the validation set, the accuracy of the cascaded GRU network is higher than that of the traditional GRU network (see Fig. 5). It shows that the cascade GRU network can identify kick cases with missing parameters more accurately. For all kick cases, the accuracy of the cascaded GRU network is 10.61% higher than that of the traditional GRU network.

The testing set includes 2,991 samples and the validation set includes 2,920 samples. There are 5,811 samples in testing and validation sets. Experimental results show that the basic GRU network identifies 4,100 of the samples. The number of calculations in the basic GRU network is 15,325,800 and that in the comprehensive GRU network is 14,909,654. So, the total number of calculations in the cascaded GRU network is 30,235,454. The number of calculations in the traditional GRU network is 49,161,060. Therefore, the cascaded GRU network reduces the computational load by 38.5% compared with the traditional GRU network.

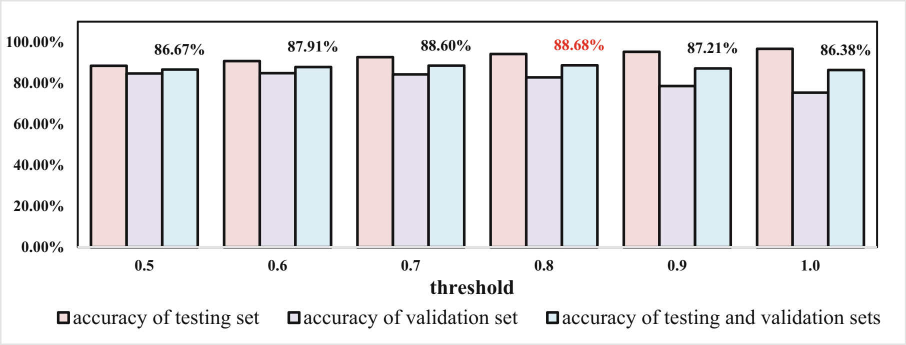


**Fig. 4.** Accuracies of testing or validation sets with different threshold.

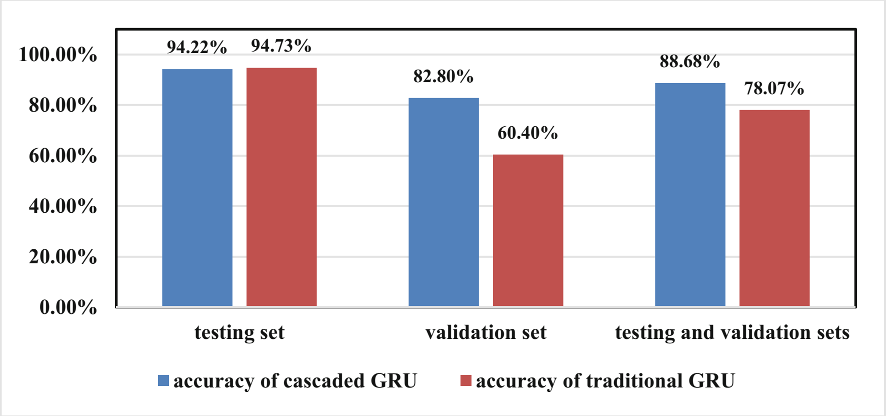


**Fig. 5.** Accuracies of cascaded GRU and traditional GRU.

## 5 Conclusion

The result of kick identification experiments using measured data shows that most kicks can be accurately identified by using pit volume and differential flow out. For drilling data without missing parameters, the kick detection model using DFO, pit volume, ROP, WOB, torque, and SPP as monitoring parameters has higher identification accuracy than the model only using pit volume and DFO. However, there are some drilling data with missing parameters collected at the field. For these data, the cascaded GRU network with adaptive monitoring parameters has higher identification accuracy than the traditional GRU network. The computational load of the cascaded GRU network is lower than the traditional GRU network. So, the applicability of intelligent kick detection methods is improved using the cascaded GRU network.

The small amount of drilling data used in this paper can only be used to verify the effectiveness of the proposed method. For practical application, it is necessary to train the model using a larger measured data set. The cascaded GRU network with adaptive monitoring parameters proposed in this paper includes two sub-networks, and only mud parameters and engineering parameters are used. If the downhole parameters are used to build a multi-level cascaded GRU network, the accuracy and robustness of the intelligent kick detection method can be further improved.

**Acknowledgments.** The project is supported by National Natural Science Foundation (Number 41974144).

## References

1. Mao, Y., Zhang, P.: An automated kick alarm system based on statistical analysis of real-time drilling data. In: Abu Dhabi International Petroleum Exhibition & Conference, Abu Dhabi, UAE (2019)
2. Tang, H., Zhang, S., Zhang, F.: Time series data analysis for automatic flow influx detection during drilling. J. Petrol. Sci. Eng. **172**, 1103–1111 (2019)
3. Tost, B.C., Rose, K., Carney, J.: Early kick detection from downhole measurements: a novel method for reducing the frequency and magnitude of loss-of-well-control events. In: Offshore Technology Conference (OTC), Houston, Texas, USA (2020)
4. Lafond, A., Leblay, F., Roguin, G.: Automated influx and loss detection system based on advanced mud flow modeling. In: SPE Annual Technical Conference and Exhibition, Calgary, Alberta, Canada (2019)
5. Blue, D., Blakey, T., Rowe, M.: Advanced mud logging: Key to safe and efficient well delivery. In: Offshore Technology Conference (OTC), Houston, Texas (2019)
6. Tarr, B.A., Ladendorf, D.W., Sanchez, D.: Next-generation kick detection during connections: influx detection at pumps stop (IDAPS) software. In: SPE Drilling & Completion vol. 31, no. 4, pp. 250–260 (2016)
7. Yalamarty, S.S., Singh, K., Kamyab, M.: Early detection of well control kick events by applying data analytics on real time drilling data. In: IADC/SPE International Drilling Conference and Exhibition, Galveston, Texas, USA (2022)
8. Shi, X., Zhou, Y., Zhao, Q.: A new method to detect influx and loss during drilling based on machine learning. In: International Petroleum Technology Conference (IPTC), Beijing, China (2019)
9. Nhat, D.M., Venkatesan, R., Khan, F.: Data-driven Bayesian network model for early kick detection in industrial drilling process. Process. Saf. Environ. Prot. **138**, 130–138 (2020)
10. Liang, H., Han, H., Ni, P.: Overflow warning and remote monitoring technology based on improved random forest. Neural Comput. Appl. **33**, 4027–4040 (2021)
11. Muojeke, S., Venkatesan, R., Khan, F.: Supervised data-driven approach to early kick detection during drilling operation. J. Petrol. Sci. Eng. **192**, 107324 (2020)
12. Yin, Q., Yang, J., Tyagi, M.: Field data analysis and risk assessment of gas kick during industrial deepwater drilling process based on supervised learning algorithm. Process. Saf. Environ. Prot. **146**, 312–328 (2021)
13. Yang, L., Han, Y., Chen, X.: Resolution adaptive networks for efficient inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2369–2378. IEEE Xplore (2020)
14. Gao, S., Huang, Y., Zhang, S.: Short-term runoff prediction with GRU and LSTM networks without requiring time step optimization during sample generation. J. Hydrol. **589**, 125188 (2020)
15. Blázquez-García, A., Conde, A., Mori, U.: A review on outlier/anomaly detection in time series data. ACM Comput. Surv. (CSUR) **54**(3), 1–33 (2021)

16. Huang, G.: Missing data filling method based on linear interpolation and lightGBM. In: Journal of Physics: Conference Series, p. 012187. IOP Publishing (2021)
17. Zhang, D., Sun, W., Dai, Y.: A hierarchical early kick detection method using a cascaded GRU network. Geoenergy Sci. Eng. **222**, 211390 (2023)
18. Gao, S., Huang, Y.Z.S.: Impact of data normalization on classification model accuracy. Res. Pap. Fac. Mater. Sci.Technol. Slovak Univ. Technol. **27**, 79–84 (2019)

# Rates Optimization of $CO_2$ Huff and Puff in Multi-stage Fracturing Horizontal Wells Based on Proximal Policy Optimization Algorithm

Rong-tao Li[1], Xiao-peng Cao[1], Zong-yang Li[1], Dong Zhang[1], Xin-wei Liao[2](✉), Qing-fu Zhang[1], and Wen-cheng Han[1]

[1] Exploration and Development Research Institute, Shengli Oilfield Company, SINOPEC, Dongying, China

[2] College of Petroleum Engineering, China University of Petroleum, Beijing, China
Xinwei@cup.edu.cn

**Abstract.** $CO_2$ huff and puff is an important replacement method for the subsequent improvement of oil recovery after the elastic development of multi-stage fracturing horizontal wells in tight reservoirs. The rates optimization of huff and puff injection and production has the advantages of low cost, easy implementation, and obvious effects. At present, the rates optimization method of huff and puff injection and production is insufficient, and the interference between different huff and puff cycles and different stages were not fully considered. This paper established a $CO_2$ huff and puff injection and production rates optimization method for multi-stage fractured horizontal wells based on the proximal policy optimization algorithm. We took the net present value as the optimization objective and huff and puff injection and production rates parameters as the optimization variables. The new method realized dynamic injection and production rates optimization with different huff and puff cycles and variable injection and production speed and variable injection and production duration, and considered the interference between various stages. The rates optimization results of $CO_2$ huff and puff in multi-stage fracturing horizontal wells indicate that the optimal project extends the backflow time by reducing the backflow rate, fully improving the utilization

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20-22 September 2023.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 225–236, 2024.
https://doi.org/10.1007/978-981-97-0272-5_19

degree of the backflow stage. At the same time, improving the efficiency of $CO_2$ injection and reducing the number of cycles significantly reduce gas injection costs, achieving optimal economic benefits and providing guidance for on-site actual $CO_2$ huff and puff injection and production rates optimization.

**Keywords:** Proximal Policy Optimization Algorithm · Multi-stage Fracturing Horizontal Wells · $CO_2$ Huff and Puff · Rates Optimization

## 1 Introduction

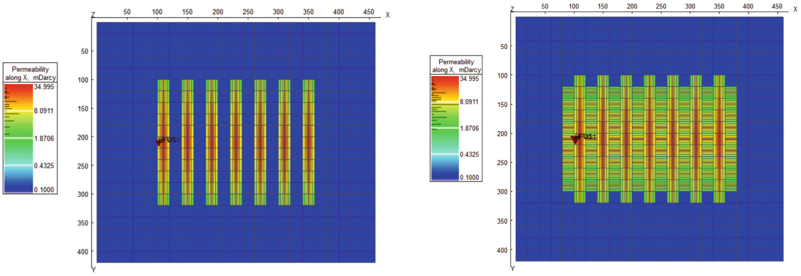
Due to the extremely low permeability of tight reservoirs, artificial fracturing is used to transform the reservoir. A large number of artificial fractures provide high-speed channels for oil and gas flow, and the single well can achieve industrial production capacity [1–3]. During the initial stage of multi-stage fracturing in horizontal wells, the oil production rate of the elastic depletion development decreases sharply [4, 5], and the $CO_2$ huff and puff development is used as the succession measure. It is difficult to form effective displacement for the $CO_2$ continuous flooding in the tight reservoir, so the $CO_2$ huff and puff development is proposed. The $CO_2$ huff and puff development effect is influenced by many factors, and it is necessary to study the impact of different influencing factors, among which the injection and production rates have a significant impact on the development effect. The injection and production rates optimization has the advantages of low cost, easy operation, and obvious effects.

At present, there is a lot of research on the influencing factors and oil displacement mechanism of the $CO_2$ huff and puff in multi-stage fracturing horizontal wells, but there is less research on the injection and production rates optimization [6–8]. The existing rates optimization methods for the $CO_2$ huff and puff have shortcomings. The optimal $CO_2$ injection rate and backflow rate are the same for different huff and puff cycle, and do not dynamically change with the change of cycle number. Moreover, the production durations of the $CO_2$ injection, soaking and production backflow have not been optimized [9]. There are significant differences in production conditions among different huff-puff cycles, and the interferences between different stages and cycles cannot be neglected, which are rarely considered in existing rates optimization methods. We established a dynamic optimization method for $CO_2$ huff and puff rates optimization in horizontal wells based on the proximal policy optimization (PPO) algorithm, achieving dynamic optimization of the huff-puff rates and time. The production interferences between different huff and puff cycles and stages have been considered, making up for the shortcomings of the existing optimization methods.

In this paper, the PPO algorithm as a gradient free optimization method has the advantages of gradient free optimization method. It does not need to obtain the partial derivative of the optimization target to the optimization variable, and it is easy to combine with commercial numerical simulation software, and achieves a wide range of applications. What's more, we deeply analyzed the mechanism of rates optimization for increasing oil recovery, and provided guidance for rates optimization on site.

## 2 Optimal Problem Description

The $CO_2$ huff and puff mechanism model for multi-stage fracturing horizontal wells consists of a geological model and a fluid model. The geological model considers the impact of fracture network morphology on the development effectiveness in Fig. 1. To demonstrate the stability of the $CO_2$ huff and puff rates optimization method based on the PPO algorithm, the geological model sets up two sets of simulation experiments with significant differences in fracture network morphology, with two wing fractures and a volume fracture network, respectively. We used local grid refinement method to simulate artificial fractures, and adopted a dual medium model to simulate matrix and natural fractures in tight reservoirs [10, 11]. Subsequent rates optimization based on the PPO requires thousands of simulation calculations. In order to significantly reduce the simulation calculation time, the mechanism model intercepts a section of a horizontal well with a length of 200m that is 7 clusters in two fractures. Using oil samples from the Chang 6 Yuan 284 block in the Ordos Basin, a fluid model was established with the component model required for reservoir numerical simulation operations.



(a) two wing fractures (b) volume fracture network

**Fig. 1.** MFHW-$CO_2$ huff and puff mechanism model

## 3 Proximal Policy Optimization Algorithm

The proximal policy optimization algorithm (PPO) is easy to achieve high solving efficiency in terms of the difference constraints between the new strategies and old strategies [12]. The PPO algorithm belong to the policy based deep reinforcement learning algorithms, which solves the problem of determining the reasonable learning step size for policy updates, so it is easy to converge and have high stability. The PPO is often treated as the preferred algorithm for Google DeepMind team to handle optimization problems.

The proximal policy optimization algorithm (PPO) has undergone complex theoretical argumentation and achieved monotonic improvement during the policy update process. The strategy has made progress every time it is updated, with the objective function of maximizing the difference between new strategies and old strategies. Every time the new strategy is updated, it performs better than the old strategy [13]. The expression for measuring the advantages and disadvantages of new and old strategies using value expectations is

$$\eta(\pi) = E_{s_0,a_0,s_1,a_1,\ldots}(\sum_{t=0}^{\infty}\gamma^t r(s_t)) \tag{1}$$

To demonstrate the monotonic improvement of the PPO policy updates, an advantage function is defined in formula (2), which represents the difference between the value of a certain action-state pair and the average value of all possible action-states under a given state. The relationship between the advantage function and the difference between the new strategies and old strategies in formula (3). When the difference between the new strategies and old strategies can be guaranteed to be non-negative, it can ensure that the new strategy monotonically improves better than the old strategy. Therefore, the optimization objective of the PPO algorithm is to maximize the difference between the old strategies and new strategies, and the optimization objective is represented by an advantage function as shown in formula (4).

$$A_\pi(s_t, a_t) = Q_\pi(s_t, a_t) - V_\pi(s_t) \tag{2}$$

$$\begin{aligned}
&\sum_{\tau\sim\pi'}(\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t, a_t))\\
&= \sum_{\tau\sim\pi'}(\sum_{t=0}^{\infty}\gamma^t (Q_\pi(s_t, a_t) - V_\pi(s_t)))\\
&= \sum_{\tau\sim\pi'}(\sum_{t=0}^{\infty}\gamma^t (r_{t+1} + \gamma V_\pi(S_{t+1}) - V_\pi(s_t)))\\
&= \eta(\pi') + E_{\tau\sim\pi'}(\sum_{t=0}^{\infty}\gamma^{t+1} V_\pi(S_{t+1}) - \sum_{t=0}^{\infty}\gamma^t V_\pi(s_t))\\
&= \eta(\pi') + E_{\tau\sim\pi'}(\sum_{t=1}^{\infty}\gamma^t V_\pi(S_t) - \sum_{t=0}^{\infty}\gamma^t V_\pi(s_t))\\
&= \eta(\pi') - E_{\tau\sim\pi'}(V_\pi(s_0))\\
&= \eta(\pi') - \eta(\pi)
\end{aligned} \tag{3}$$

where, $\tau$ is the sampling trajectory that is determined by state probability, transition probability, and policy strategy.

$$\begin{aligned}
J^\theta(\theta) = \eta(\pi') \text{ - } \eta(\pi) &= \sum_{\tau\sim\pi'}(\sum_{t=0}^{\infty}\gamma^t A_\pi(s, a))\\
&= \sum_{t=0}^{\infty}\sum_{s}P(s_t = s|\pi')\sum_{a}\pi'(a|\mathrm{s})\gamma^t A_\pi(s, a)
\end{aligned}$$

$$=\sum_{s}\sum_{t=0}^{\infty}\gamma^{t}P(s_t=s|\pi')\sum_{a}\pi'(a|\mathrm{s})A_{\pi}(s,a)$$

$$=\sum_{s}\rho_{\pi'}(s)\sum_{a}\pi'(a|\mathrm{s})A_{\pi}(s,a) \tag{4}$$

where, $\rho_{\pi}(s)$ represents the sum of state probability for any time step state

$$\rho_{\pi}(s)=P(s_0=s)+\gamma P(s_1=s)+\gamma^2 P(s_2=s)+\ldots=\sum_{t=0}^{\infty}\gamma^{t}P(s_t=s) \tag{5}$$

The proximal policy optimization algorithm (PPO) aims to improve sample utilization by transforming on-line learning of strategies into off-line learning based on the importance sampling theorem. Samples generated from historical old strategies can be repeatedly sampled for learning, resulting in a significant improvement in sample utilization. The cumulative reward expectation expression is transformed into formula (6) after importance sampling. By introducing importance weight correction, the cumulative reward expectation of the new strategy distribution can be calculated using the old strategy distribution, but the difference between the two action probability distributions of the new strategy and the old strategy cannot be too large. Similarly, the PPO optimization target is transformed by importance sampling as shown in Formula (7).

$$E_{x\sim p}(f(x))=\int f(x)p(x)dx=\int f(x)\frac{p(x)}{q(x)}q(x)dx=E_{x\sim q}(f(x)\frac{p(x)}{q(x)}) \tag{6}$$

$$J^{\theta'}(\theta)=E_{(s,a)\sim\pi_{\theta'}}(A_{\pi_{\theta'}}(s,a)\frac{\pi_{\theta}(s,a)}{\pi_{\theta'}(s,a)}) \tag{7}$$

The PPO algorithm achieves monotonic improvement in policy updates by ensuring that the expected difference between the new strategies and old strategies is not negative. Therefore, the PPO algorithm is set to maximize the expected difference between the new strategies and old strategies, and the expected difference between the new and old strategies is expressed through an advantage function for easy calculation and implementation. In order to further improve the sample utilization rate, the importance sampling theorem is introduced, and the old strategy distribution replaces the new strategy distribution. From online learning to offline learning, the samples generated by the old strategy in history are repeatedly sampled. The sample utilization rate is significantly improved, but the distribution of the new strategies and old strategies cannot be too different, and certain constraints need to be met.

The PPO algorithm improves the constraint on the difference between new and old strategies, and uses the Clip truncation function to truncate the importance ratio of the new strategies and old strategies, limiting its value range and ensuring that the fluctuation amplitude of each gradient update is reasonable and avoiding excessive fluctuation, as shown in formula (8). What's more, using the minimum value function ensures that all minimum values can achieve excellent performance.

$$J^{\theta'}(\theta)=E_{(s,a)\sim\pi_{\theta'}}(\min(A_{\pi_{\theta'}}(s,a)\frac{\pi_{\theta}(s,a)}{\pi_{\theta'}(s,a)},clip(A_{\pi_{\theta'}}(s,a)(\frac{\pi_{\theta}(s,a)}{\pi_{\theta'}(s,a)},1-\varepsilon,1+\varepsilon)))$$

$$\approx \sum_{(s,a)} \min(A_{\pi_{\theta'}}(s,a)\frac{\pi_\theta(s,a)}{\pi_{\theta'}(s,a)}, clip(A_{\pi_{\theta'}}(s,a)(\frac{\pi_\theta(s,a)}{\pi_{\theta'}(s,a)}, 1-\varepsilon, 1+\varepsilon)) \tag{8}$$

## 4 CO2 Huff-n-Puff Rates Optimization Based on the PPO

The rates optimization of $CO_2$ huff and puff in multi-stage fracturing horizontal wells (MFHW) based on PPO algorithm is an optimization problem, which consists of optimization objective, optimization decisions, and constraint conditions.

### 4.1 Optimization Objective

The determination of optimization objective requires comprehensive consideration of the physical problem characteristics of $CO_2$ huff and puff rates optimization and the principle of the PPO algorithm. If the optimization target is cumulative oil production or oil recovery, without considering the impact of gas injection cost, the intelligent agent achieves maximum oil recovery by significantly increasing the cumulative gas injection amount resulting in too much high cost. Therefore, we take net present value (NPV) as the optimization objective in Formula (9), and pursue the maximum oil recovery while taking into account the impact of $CO_2$ injection cost, and obtain the maximum economic benefits.

$$NPV = \sum_{n=1}^{N} \frac{P_o Q_o^{\Delta t^n} - C_{CO_2-INJ} Q_{CO_2-INJ}^{\Delta t^n} - C_{CO_2-PRO} Q_{CO_2-PRO}^{\Delta t^n}}{(1+b)^{t^n/365}} \tag{9}$$

where, $P_o$ is crude oil price, yuan/ton; $C_{CO_2-INJ}$ and $C_{CO_2-PRO}$ are the prices for injecting $CO_2$ and produced $CO_2$ treatment, respectively, yuan/ton. $Q_o^{\Delta t^n}$, $Q_{CO_2-INJ}^{\Delta t^n}$ and $Q_{CO_2-PRO}^{\Delta t^n}$ are the cumulative oil production, $CO_2$ injection, and $CO_2$ production in the n-th injection production adjustment time step, tons; $N$ is the total number of rates optimization adjustment time steps; $\Delta t^n$ is the n-th adjusted time step, in days; $b$ is the annual interest rate, %.

### 4.2 Optimization Decision

According to the principle of deep reinforcement learning PPO algorithm, it is necessary to determine the action space variables and state space variables, where the action space variables are the optimization decisions. There are five action space variables for the rates optimization of the $CO_2$ huff and puff in multi-stage fracturing horizontal wells based on the PPO algorithm, namely gas injection rate, gas injection time, soaking time, backflow rate, and backflow time. The maximum variation range of rates variables in adjacent adjustment time steps is within 20%, to prevent the variables variation range from being too large or too small, resulting in difficulties in reservoir simulation convergence and on-site construction. State space variables are needed to clearly describe the huff and puff production characteristics, providing information for intelligent agents to select the optimal injection and production action based on specific environmental states. We selected $CO_2$ huff and puff well production data as a total of 9 state space variables, which are time steps, daily gas injection, daily gas production, daily oil production, production gas oil ratio, bottom hole flow pressure, cumulative gas injection, cumulative gas production, and cumulative oil production. The state space variables should be as comprehensive as possible to improve the accuracy of describing the state of huff and puff injection and production environment.

### 4.3 Constraint Condition

The constraints of the physical problem of $CO_2$ huff and puff injection production optimization for multi-stage fracturing horizontal wells are: the upper limit of bottom hole flowing pressure in the injection stage that is lower than 90% of the rock breakdown pressure. In the backflow stage, the lower limit of the bottom hole flow pressure of the production well is higher than 10MPa. On the one hand, the lower the bottom hole flow pressure of in the backflow stage, the greater the production pressure difference provided, and the higher the reservoir utilization degree. On the other hand, setting a reasonable lower limit of bottom hole flow pressure in the backflow stage can control formation damage within a certain range.

### 4.4 Optimization Method Process

The optimal strategy for $CO_2$ rates optimization in multi-stage fracturing horizontal wells is achieved through continuous interaction between the PPO intelligent agent and the injection and production environment. At each rates optimal adjustment time step, the agent transfers the action variable injection production rates and time steps to the reservoir numerical simulation environment. With the running calculation of the reservoir simulator, the agent obtains the simulation results of the well production data, which are used to calculate the immediate NPV of this adjustment time step and form the state space variable of the next adjustment time step, and were stored as a sample in the experience pool. In the subsequent reservoir simulation round calculation, the target action selection network searches the historical experience pool for the optimal action corresponding to the maximum cumulative reward. The online action selection network adjusts the weight in the direction of maximizing the cumulative reward. With the increase of reservoir simulation rounds, the optimization target cumulative NPV gradually converges to the maximum value, and the optimal huff and puff strategy were achieved. The method process of the $CO_2$ huff and puff rates optimization in multi-stage fracturing horizontal wells is shown in Fig. 2.

## 5 Algorithm Implication and Simulation Results

We built the huff-n-puff rates optimization method of the multi-stage fracturing horizontal well based on the PPO algorithm. We applied the rates optimization method to two group simulation cases to demonstrate the effectiveness and stability of the new rates optimization method. What's more, we further analyzed the reasons behind the phenomenon of the optimal case, and provided guidance for rates optimization on site.

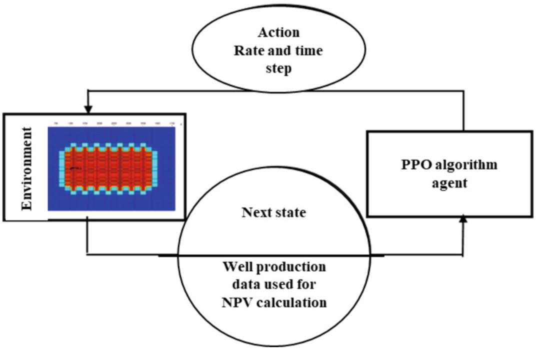


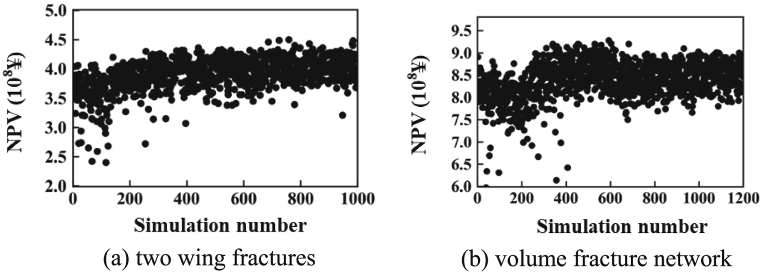
**Fig. 2.** Rates optimization method process for MFHW-$CO_2$ huff and puff

### 5.1 Demonstrate Optimization Method Effectiveness and Stability

Two groups simulations of the $CO_2$ huff and puff rates optimization in multi-stage fracturing wells with two wing fractures and volume fractures were set. If the NPV of the optimal case is significantly higher than the base case, it indicates the effectiveness of the new rates optimization method. If two sets of simulations with significant differences in main influencing factors can converge to the optimal solution, it indicates that the rates optimization method has high stability and wide applicability. Otherwise, it is only applicable to specific situations.

We applied the new rates optimization method to the two group simulation cases. The economic parameters used to calculate the NPV are as follows: the crude oil price is 2818 yuan/ton, the injected $CO_2$ price is 550 yuan/ton, the produced $CO_2$ treatment cost is 30 yuan/ton, and the annual benchmark interest rate is 8%. The PPO algorithm is based on the Actor-Critical algorithm framework and consists of an action selection Actor network and a value evaluation Critical network. The intelligent agent selects specific actions through the action selection network based on the state of the injection and production environment. The Actor network consists of three layers, with 64 neurons in the middle layer, 9 neurons in the input layer and 2 neurons in the output layer. The intelligent agent evaluates the value of specific states and action space through the value evaluation network. The Critical network consists of three layers, with 64 neurons in the middle layer, 9 neurons in the agent state space dimension in the input layer, and 1 neuron in the target value dimension in the output layer.

As Fig. 3 shows the optimal target cumulative NPV gradually converges to the optimal value around the 500th simulation round and remains near the optimal value until the end of the 1000th simulation round. We took the P5 optimal scheme that is higher than 95% optimal scheme as the optimal case, rather than the highest optimal scheme to increase credibility. The optimal target NPV of the base case and the P5 optimal case in horizontal well with two wing fractures are $3.80 \times 10^8$ ¥ and $4.28 \times 10^8$ ¥, respectively. The NPV of the base case and the P5 optimal case in horizontal well with volume fracture network are $8.32 \times 10^8$ ¥ and $8.90 \times 10^8$ ¥ respectively. The NPV of the P5 optimal case of the two wing fractures and the volume fracture network are 12.63% and 6.97% higher than the corresponding base case respectively, which proves the effectiveness of the new rates optimization method based on the PPO algorithm. The optimal cases with significant differences in main influencing factors fracture network morphology can converge to the optimal value, indicating that the established rates optimization method has high stability and wide application range.



(a) two wing fractures (b) volume fracture network

**Fig. 3.** Convergence change of target NPV of MFHW-$CO_2$ huff and puff

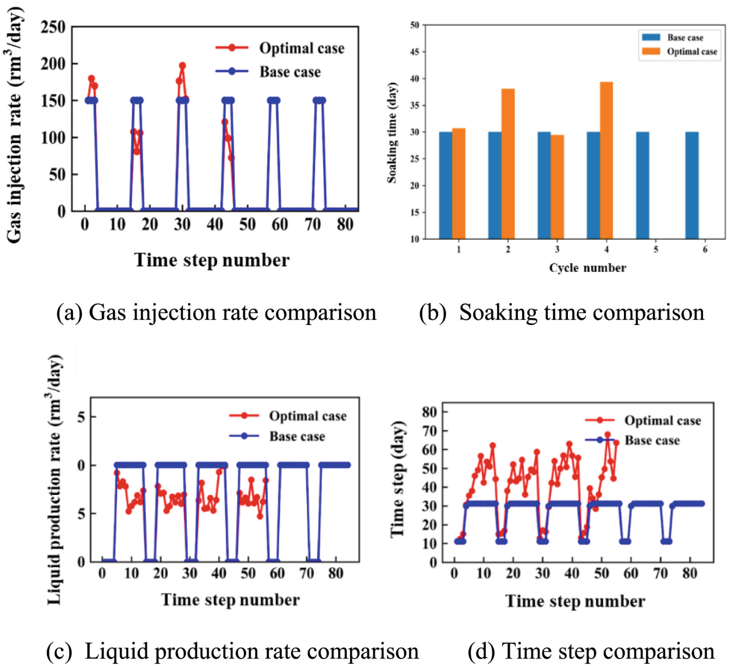
## 5.2 Analyze Rates Optimization Case Oil Increasing Mechanism

By comparing and analyzing the rates optimal case with the base case, and taking the horizontal well with volume fracture network as an example, the oil increasing mechanism was analyzed from different perspectives, providing theoretical guidance for on-site operation.

The rates optimization system obtained from $CO_2$ huff and puff in multi-stage fracturing horizontal wells, including the optimal gas injection rate, gas injection time, soaking time, backflow rate, and backflow time, are shown in Fig. 4, respectively. We firstly analyzed the reasons why the optimal target NPV of the optimal case is higher than that of the base case. By comparing and analyzing the cumulative production data as shown in Fig. 5, it can be concluded that the main difference lies in the cumulative gas injection volume and cumulative gas production volume. The optimal case has significantly lower cumulative gas injection volume and cumulative gas production volume than the base case. The cumulative oil production of the optimal case is slightly higher than that of the base case. Thus, the optimal project achieves the optimal target NPV by reducing the gas injection cost results from the decreased cumulative gas injection, and retains the revenue for cumulative oil production slightly change.

From the aspect of different production stages of injection, soaking and backflow, we analyzed the oil increasing mechanism of $CO_2$ huff-n-puff in multi-stage fracturing horizontal wells. The injected $CO_2$ during the injection phase increases formation pressure, and reduces formation damage caused by pressure sensitivity. During the soaking stage the injected $CO_2$ dissolves and diffuses in the crude oil. In the backflow production stage, the immiscible flooding and the dissolved gas flooding are the main mechanisms for the increased oil production.

As the Fig. 4(a) shows, the optimal case reduces the cycle number from 6 to 4. The gas injection rate of the optimal case was significantly lower than that of the base case in the second the fourth cycles, while was higher than the base case in the first and



(a) Gas injection rate comparison (b) Soaking time comparison

(c) Liquid production rate comparison (d) Time step comparison

**Fig. 4.** Rates optimization system of the MFHW-$CO_2$ huff and puff

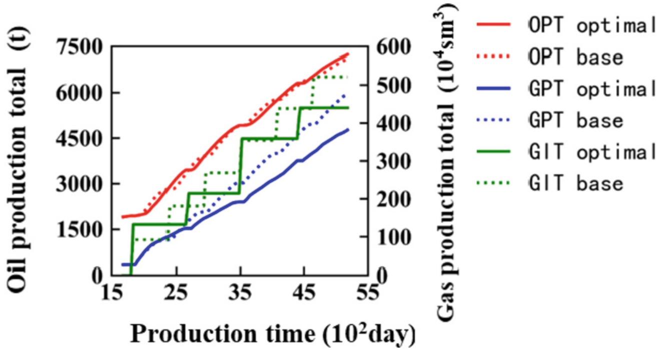


**Fig. 5.** Cumulative production data comparison

third cycles. From the aspect of average formation pressure, the optimal case for the first and third huff and puff cycles of the volume fracture network are higher than the base case, while other cycles are similar to the base case. During the injection phase, the larger the cumulative gas injection in a single huff and puff cycle, the greater the increase in average formation pressure, and the smaller formation damage caused by the stress sensitivity. Due to the small difference in the average pressure between the base case and optimal case, the degree of reservoir damage is slight different. However, the increased the cumulative gas injection in a single cycle leads to high cost and poor economic benefits.

During the soaking stage the injected $CO_2$ interacts with crude oil in the reservoir matrix through molecular diffusion. There is a slight difference in the soaking time between the base case and the optimal case, as shown in Fig. 4(b). The recovery rate of $CO_2$ huff and puff in multi-stage fracturing horizontal well increases with the increased soaking time. However, the influence degree of soaking time on the development effect of $CO_2$ huff and puff is very small and can be neglected.

Comparing the liquid production rate and time steps during the backflow stage, the optimal case has significantly lower backflow speed than the base case, while the reverse flow time steps were significantly higher than the base case, as shown in Fig. 4(c) and Fig. 4(d). It indicates that the backflow stage of the optimal case makes full use of formation energy and has a higher utilization degree, which makes the main contribution to the increased oil recovery.

Overall, the oil recovery of the optimal case and the base case are nearly equal, because the adverse effects of reduced cumulative gas injection and formation damage caused by pressure sensitive for the optimal case are offset by the favorable effects of reducing the backflow rate and increasing the backflow time, making the backflow stage more fully utilized.

## 6 Conclusions

(1) A new $CO_2$ huff and puff rates optimization method for horizontal wells with multi-stage fracturing based on the PPO algorithm has been established, and the effectiveness and stability have been demonstrated.

(2) The optimal case makes the backflow stage more fully utilized by reducing the backflow rate and increasing the backflow time, and also reduces cumulative $CO_2$ injection, achieving high economic benefit.

## References

1. Zhang, J., Bi, H., Xu, H., et al.: New progress in tight oil exploration and development abroad and its reference significance. J. Pet. **36**(2), 127–137 (2015)
2. Jiang, Z., Zhang, W., Liang, C., et al.: Basic characteristics and evaluation factors of shale oil reservoirs. J. Pet. **35**(1), 184–196 (2014)
3. Jia, C., Zou, C., Li, J., et al.: Main types, basic characteristics, and resource prospects of China's tight oil evaluation standards. J. Pet. **33**(3), 343–350 (2012)
4. Wei, Y., Ran, Q., Tong, M., et al.: Full cycle productivity prediction model for tight oil fracturing horizontal wells. J. Southwest Pet. Univ. (Nat. Sci. Ed.) **38**(1), 99–106 (2016)
5. Agboada, D.K., Ahmadi, M.: Production decline and numerical simulation model analysis of the Eagle Ford shale oil play. In: SPE Western Regional & AAPG Pacific Section Meeting Joint Technical Conference. OnePetro (2013)
6. Sun, J., Zou, A., Sotelo, E., et al.: Numerical simulation of CO2 huff-n-puff in complex fracture networks of unconventional liquid reservoirs. J. Nat. Gas Sci. Eng. **31**(5), 481–492 (2016)

7. Ding, M., Gao, M., Wang, Y., et al.: Experimental study on CO2-EOR in fractured reservoirs: Influence of fracture density, miscibility and production scheme. J. Petrol. Sci. Eng. **17**(4), 476–485 (2019)
8. Zuloaga, P., Yu, W., Miao, J., et al.: Performance evaluation of CO2 Huff-n-Puff and continuous CO2 injection in tight oil reservoirs. Energy **13**(4), 181–192 (2017)
9. Sanchez-Rivera, D., Mohanty, K., Balhoff, M.: Reservoir simulation and optimization of Huff-and-Puff operations in the Bakken Shale. Fuel **14**(7), 82–94 (2015)
10. Du, C.M., Zhang, X., Zhan, L., et al.: Modeling hydraulic fracturing induced fracture networks in shale gas reservoirs as a dual porosity system. In: International Oil and Gas Conference, China (2010)
11. Weng, X., Kresse, O., Cohen, C., et al.: Modeling of hydraulic-fracture-network propagation in a naturally fractured formation. SPE Prod. Oper. **26**(4), 368–380 (2011)
12. Schulman, J., Wolski, F., Dhariwal, P., et al.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
13. Schulman, J., Levine, S., Abbeel, P., et al.: Trust region policy optimization. In: International Conference on Machine Learning. PMLR (2015)

# Application of Artificial Intelligence Fracture Detection in Hechuan Area

Zhe Wang(✉), Weili Hou, Huitian Lan, Tingting Qiao, Shan Wang, and Shuang Han

Exploration and Development Research Institute of Daqing Oilfield, Daqing, China
Wangzhe87@petrochina.com.cn

**Abstract.** The development of strike slip faults in the central part of the Sichuan Basin is influenced by the structure, and the high yield wells reveal that strike slip faults have a close relationship with reservoir control. This article uses 3D seismic data from the HC125 work area in the Hechuan Tongnan area of the Sichuan Basin to carry out identification of strike slip faults based on artificial intelligence. Firstly, preprocess seismic data to improve the imaging characteristics of strike slip faults in seismic profiles. Secondly, developing method for edge coherence enhancement to highlight the faults boundaries. Finally, U-Net convolutional neural network machine learning method is used to identify main faults, and disorder detection technology is used to identify associated fractures and small-scale faults. According to this, a comprehensive detection technology suitable for strike-slip faults in the central Sichuan region will be formed. Compared with conventional fracture detection technology, artificial intelligence technology for fracture detection has a relatively high fault resolution, and the continuity and interpretability of deep fracture have been greatly improved; Results of multi-scale Fault Detection guarantees the research on the Mechanism of Controlling hydrocarbon accumulation through Strike-slip Faults in the Central Sichuan Basin. This technology effectively improves the interpretation accuracy and classification accuracy of strike slip faults in the central Sichuan region.

**Keywords:** Strike-slip fault · Deep learning · Central Sichuan region · Data preprocessing

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20–22 September 2023.

This paper was selected for presentation by the IFEDC Committee following review of information contained in an abstract submitted by the author(s). Contents of the paper, as presented, have not been reviewed by the IFEDC Technical Team and are subject to correction by the author(s). The material does not necessarily reflect any position of the IFEDC Technical Committee its members. Papers presented at the Conference are subject to publication review by Professional Team of IFEDC Technical Committee. Electronic reproduction, distribution, or storage of any part of this paper for commercial purposes without the written consent of IFEDC Organizing Committee is prohibited. Permission to reproduce in print is restricted to an abstract of not more than 300 words; illustrations may not be copied. The abstract must contain conspicuous acknowledgment of IFEDC. Contact email: pa-per@ifedc.org.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 237–244, 2024.
https://doi.org/10.1007/978-981-97-0272-5_20

## 1 Introduction

In recent years, exploration and geological studies have made it clear that the Sichuan Basin as a whole is a complex oil-bearing system, and strike-slip faults play an important role in controlling multi-layer complex petroliferous accumulation [1]. Therefore, it is very important to accurately characterize the strike-slip faults. Fault plays an important controlling role in improving reservoir physical property and hydrocarbon migration and accumulation [2]. Conventional fracture detection methods mainly include coherence body technology, curvature attribute identification technology, variance body technology, ant body tracking technology and so on [3]. With the continuous development of technology, the "three-low phenomenon" of low precision, low continuity and low resolution of conventional fault identification has seriously restricted the complex strike-slip fault interpretation [2, 4]. In recent years, with the development of artificial intelligence technology, the geophysical field has been trying to use artificial intelligence technology to serve for seismic exploration. Scholars in the field of geophysics begin to explore the application of machine learning methods to fault identification. Han Chengyang [5] et al. proposed to use CNN model to predict faults, and the predicted results are roughly consistent with the artificial interpretation results, but its accuracy and resolution still need to be improved. Zhang Li [6] et al. proposed strike-slip fault identification based on full convolutional neural network technology, compared U-Net and SegNet two full convolutional algorithms for optimization, and used the construction-oriented method to improve network performance and enhance the generalization ability of network structure. Yang Wuyang [7] et al. proposed that U-Net and Res-50 residual modules jointly build a new network ResU-Net to improve the noise resistance of the model, strengthen its generalization ability, and improve the accuracy of fault prediction. Chang Dekuan [8] et al. proposed the combination of ResNet deep residual network and U-Net architecture to characterize multi-scale fault information. A large number of studies have shown that the artificial intelligence technology has better accuracy, resolution and continuity than the traditional coherence and curvature attributes in fault identification. In this paper, the fracture detection algorithm based on U-Net convolutional neural network image segmentation is adopted, and the large-scale trunk fractures are identified by machine learning. At the same time, the messy detection of multi-frequency coherence is adopted to identify the associated fractures and cracks, so as to form a set of comprehensive detection methods for different fracture levels in Hechuan area.

## 2 Geological Overview

The Sichuan Basin is located in the northwest of the Yangtze Block and has experienced multi-stage tectonic-sedimentary cycles with complex structure. Meanwhile, a large number of studies have proved that the strike-slip fault has a significant control over the reservoir transformation, oil and gas migration and accumulation. The strike-slip structure in the central Sichuan area developed on the early basement. They can be divided into several stages of evolution, such as the embryonic stage of Nanhua strike-slip fault, the Late Sinian-Early Cambrian right-lateral tension-torsion strike-slip movement, the Ordovician--Permian pre-weak compression and tension-torsion strike-slip movement, and the late Permian left-lateral weak tension-torsion strike-slip movement [9].

The Tongnan block of Hechuan is located in the central Sichuan region, where karst fracturevugg-type reservoirs are developed under the influence of Caledonian movement and Yunnan movement uplift before Permian and Dongwu movement uplift in Permian [10]. The strike-slip faults developed at the same time are the main controlling factors for the formation of efficient reservoirs, so it is of great geological significance for the identification of strike-slip faults.

# 3 Technical Principle and Process

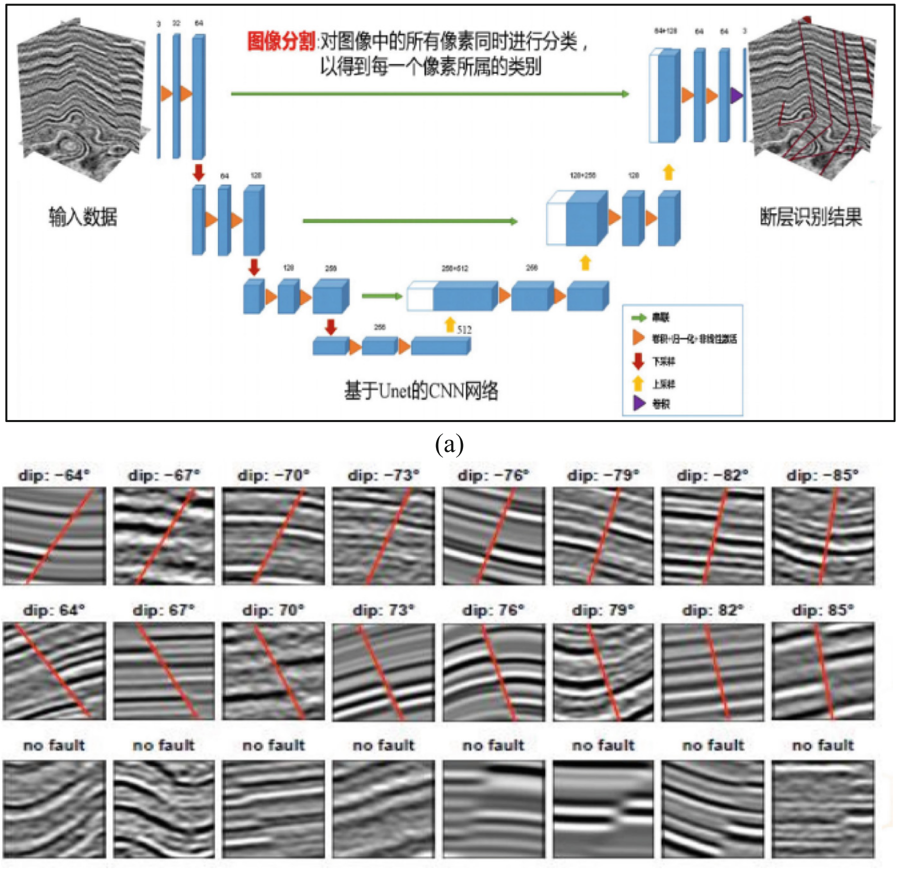
In view of the strike-slip fault movement of the Permian system with weak pressure and torsion, the conventional fault identification technology is difficult to identify the strike-slip fault and the associated small faults. The strike-slip faults in central Sichuan are mainly developed in NNW direction. A set of comprehensive fault detection techniques for this area is proposed.

## 3.1 Seismic Data Preprocessing

Azimuth seismic data perpendicular to the strike of the fault is selected, denoising is optimized for the seismic data, and fault enhancement processing is adopted for the seismic data to highlight characteristics of the fault.

## 3.2 Deep Learning Fault Detection Technology Based on U-Net Convolutional Neural Network

Compared with conventional seismic interpretation methods such as coherence and curvature, deep learning can accurately reflect the characteristics of main fault in seismic data with its powerful calculation and learning ability. In this paper, the convolutional neural network (CNN) technology based on U-Net architecture is used to treat the fault as an image segmentation problem, which is mainly composed of encoder and decoder. The encoder consists of a convolutional layer and a pooling layer. The image after convolution is nonlinear activated by using ReLU function. During downsampling, the number of activated feature image channels is increased by two times and the image size is reduced by two times. In the encoder part is the deconvolution of the feature image, enlarging the size of the feature image twice. The encoder and decoder are connected in the form of full connection layer, and the parameters are constantly adjusted based on ReLU activation function. In the output layer, 1x1 convolution is used to check the feature image convolution restoration to generate the original image, and finally, the fault is labeled by Sigmoid function [7]. Technical features and advantages of this method: training data are obtained by establishing a large number of different fracture modes and synthesizing seismic models, and given the "fault" label, to achieve the supervised deep learning based on big data, and the U-Net based convolutional neural network has a strong generalization ability in fault classification [4] (see Fig. 1).



(a)

(b)

**Fig. 1.** U-Net network structure diagram (a) and fault sample setup diagram (b).

### 3.3 Messy Fracture Detection Based on Amplitude Gradient Vector Coherence

Carry out the identification of secondary faults, small faults, fractures and other small and medium scale faults. The amplitude gradient vector messy fault detection technology is based on the third generation of coherent method to implement fault detection, the core idea of this method is (assuming that the fault plane is a plane in a local area, search the messy degree of seismic amplitude gradient vector through each azimuth and dip Angle in three-dimensional space, find the most messy degree of the surface indicating the fault location, According to this, the fault body optimization is carried out, and finally the fault messy body reflecting the characteristics of the section is obtained [11]. Search the messiness of the amplitude vector along a direction v of a seismic sample point s, and construct the gradient construction tensor field T($s$, $v$), whose expression (1) is as follows:

$$\mathrm{T}(s, v) = \begin{bmatrix} \int D_x^2 SW_N(s, v) & \int D_x D_y W_N(s, v) & \int D_x D_t W_N(s, v) \\ \int D_x D_y SW_N(s, v) & \int D_y^2 SW_N(s, v) & \int D_y D_t SW_N(s, v) \\ \int D_x D_t SW_N(s, v) & \int D_y D_t SW_N(s, v) & \int D_t^2 SW_N(s, v) \end{bmatrix} \tag{1}$$

where Dx, Dy and Dt are the change rate of seismic amplitude along x, y, z over time t respectively, $SW_N$(s, $v$) are the smoothing factor along azimuth v, and the smoothing function is the multi-point Gaussian function. After establishing the matrix in the direction of s, Eq. (2) is used to obtain the disorder property of the amplitude vector, where are the first, second and third eigenvalues respectively $\lambda_1, \lambda_2, \lambda_{3.}$ This method makes the coherent data more accurately indicate the fault information, the resolution is higher than the conventional coherent data and the fault information is richer.

$$\mathrm{F}(s, v) = \frac{3}{2} \frac{\lambda_2(s, v) + \lambda_3(s, v)}{\lambda_1(s, v) + \lambda_2(s, v) + \lambda_3(s, v)} \tag{2}$$

## 4 Application

The study area is located in the middle and low moderate tectonic belt of Sichuan. Hechuan area was influenced by Caledonian, Hercynian and Indochinese tectonic movements. The top surface of Maokou Formation developed three stages of near-east-west, northeast and northwest strike-slip faults, and the HC125 working area in this study mainly developed near-east-west and northwest strike-slip faults. Under the influence of Dongwu Movement, the top of Maokou Formation is generally affected by tectonic denudation and regional unconformity, which is generally affected by palaeogeomorphic elevation difference and erosion. At the same time, the development of strike-slip faults is conducive to the infiltration of surface water into the fracture zone, which is conducive to the formation and development of karst reservoirs.

### 4.1 Preprocessing of Seismic Data

Denoising and fault enhancement processing were carried out on the seismic data, and the comparison between the original seismic data and the pre-processed data showed that the edge of deep and large faults was clearer, local hidden faults were prominent, and the fracture zone was more clearly delineated, which provided a reliable data basis for fault identification (see Fig. 2.a, Fig. 2.b).

### 4.2 Main Fault Identification

Figure 4 shows the composite diagram of fault probability body and seismic profile obtained by two methods of maximum likelihood attribute and deep learning. The maximum likelihood body is chaotic and disorderly, and the continuity of deep fault is unclear. Deep learning has high resolution and good continuity for deep fault (see Fig. 2 c, Fig. 2 d).

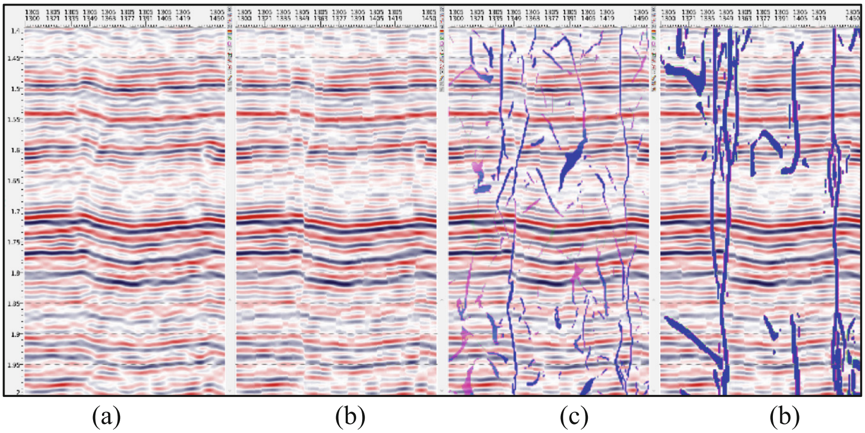


**Fig. 2.** Seismic profile of line1305 before (a) and after preprocessing (b), Maximum Likelihood overlay profile (a), AI probability volume and seismic data overlay profile (b)

### 4.3 Messy Attribute Detection of Associated Fractures

Figure 3 shows that the messy detection profile has richer fracture information than the deep learning detection profile, and the fracture zone and secondary fracture associated with the main fracture are described more clearly.

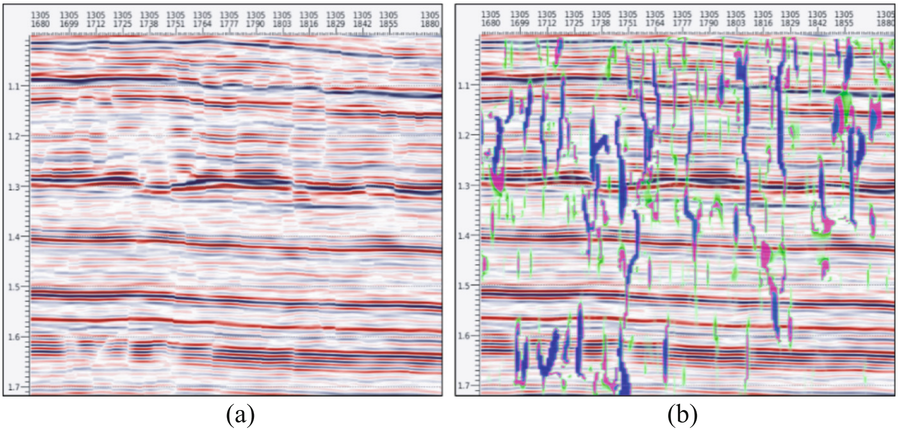


**Fig. 3.** Seismic profile (a) and seismic composite profile of messy body (b)

From the above analysis, it can be seen that the deep learning is clear in characterizing the main faults, and the fault distribution characteristics are consistent with the geological understanding. Since the ability of characterizing minor faults and other secondary associated faults is relatively weak, the characterization of minor faults and cracks is enhanced with the messy detection technology, so as to obtain the development characteristics and distribution rules of multi-scale faults in this region. Relative to the coherence, curvature plan and multi-scale fault overlap, coherence plan resolution is low and curvature diagram fault classification is not obvious (Fig. 4).

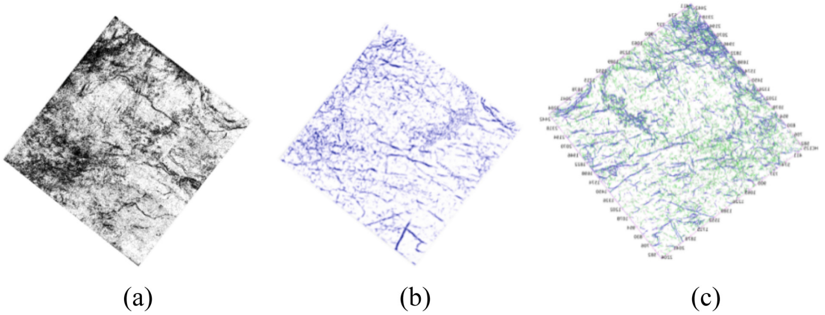


**Fig. 4.** Coherence slice (a), curvature slice (b) and fracture composite slice (c) of HC125 working area

The TS10HC well deployed in the working area used the technique for fault detection, predicting that a small fault (The position indicated by the yellow arrow in Fig. 5.b) would be drilled in the deviation section. Small faults and fractures are developed in the northwest direction near well TS10HC, and the strike of large faults is consistent with geological understanding, with rich details of small faults and fractures (see Fig. 5).

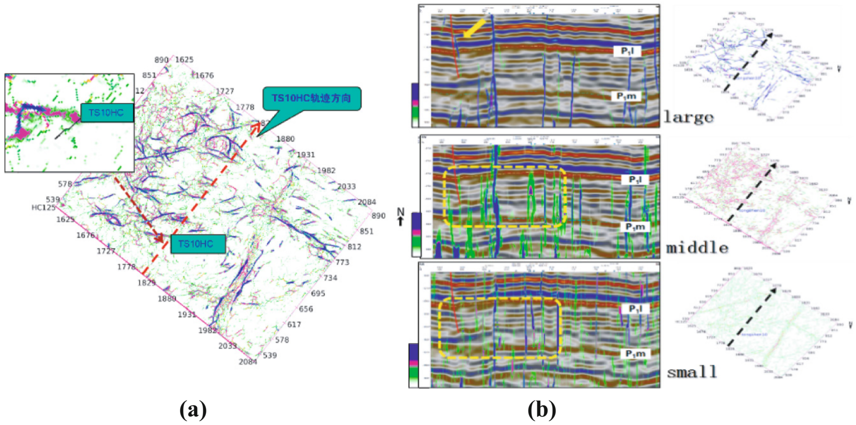


**Fig. 5.** Slice of Multi-scale fault superposition in the TS10HC well area (a), Seismic superimposed profiles of faults at different scales

## 5 Application

1. Messy detection technology is suitable for identifying the faults associated with strike-slip faults and local hidden small faults. The combination of deep learning and messy technology provides a guarantee for the accurate characterization of the fault system.

2. Compared with conventional fracture recognition technology, deep learning fracture detection based on U-Net convolutional neural network algorithm has a clearer and higher resolution. It can well reflect the distribution space of faults and accurately predict the distribution law of strike-slip faults in Hechuan area, which is highly consistent with the geological understanding, proving the feasibility and reliability of artificial intelligence technology in the field of fault detection.
3. After applying this technology in the research area, a new TS10HC was drilled for fault-controlled karst reservoir in the area. The strike-slip fault next to the well is accurately predicted. At the same time, the small fault associated with the strike-slip fault is predicted in the inclined section. The leakage occurs there during the drilling process, which proves the accuracy of the fault prediction of this technology, and further proves that the technology can be extended in the detection of the strike-slip fault in the Hexhuan area (see Fig. 5).

## References

1. Guan, S., Jiang, H., Lu, X., et al.: Strike-slip fault system and its controlling effect on oil and gas in central Sichuan Basin. Acta Petrolei Sinica **43**(11), 1542–1557 (2022)
2. Debo, M., Yimin, Z., Yintao, Z., et al.: Application of maximum likelihood attribute to fault identification: a case study on identification of Ordovician strike-slip faults in Rewapu block, Halahatang area, Tarim Basin. Nat. Gas Geosci. **29**(6), 817–825 (2018)
3. Li, T., Hou, S., Ma, S., et al.: Review of fault identification methods and research progress. Prog. Geophys. **33**(4), 1507–1514 (2018)
4. Junan, C., Haidong, C., Wei, G., et al.: An application example of fault detection based on deep learning and edge enhancement. Oil Geophys. Prospect. **57**(6), 1304–1316 (2022)
5. Han, C., Zou, G., Li, Z.: Analysis of automatic seismic fault identification based on Convolutional neural network. In: Proceedings of the National Conference on Mineral Exploration (2021)
6. Zhang, L., Lv, F., Shang, K., et al.: Strike-slip fault identification technique based on full convolutional neural network: a case study of Toputai area. J. Yangtze Univ. 19(4), 38–48 (2022)
7. Yang, W., Yang, J., Chen, S., et al.: Fault detection of seismic data based on U-Net deep learning network. Oil Geophys. Prospect. **56**(4), 688–697 (2021)
8. Chang, D., Yong, X., Wang, Y., et al.: Fault recognition method of seismic data based on deep convolutional neural network. Oil Geophys. Prospect. **56**(1), 1–8 (2021)
9. Ma, B., Liang, H., Wu, G., et al.: Formation and evolution of multi-stage strike-slip faults in central Sichuan Basin. Petrol. Explor. Dev. **50**(2), 333–345 (2023)
10. Yin, C., Shi, J., Zheng, J., et al.: Petroleum geology and oilfield development in daqing. **42**(1), 1–10 (2023)
11. Zheng, M., Chen, K., Cai, J., et al.: Shale gas fracture prediction based on amplitude gradient disorder detection algorithm in Changning area. Prog. Geophys. **37**(5), 21010–2117 (2022)

# Hybrid Q&A Method for Knowledge Graph and Documents of Global Petroliferous Basins

Ting-yu Ji[1], Da-wei Li[2](✉), Ming-cai Yuan[1], Min Niu[2], Shi-yun Mi[2], Xiao-yu An[2], Fen Wang[1], and Qiang Lu[1]

[1] China University of Petroleum (Beijing), Beijing 102249, China
[2] PetroChina Research Institute of Petroleum Exploration and Development, Beijing 100083, China
leedw@petrochina.com.cn

**Abstract.** The massive data and information in the petroliferous basins formed by exploration and development are extremely valuable. Thus, they need to be deeply mined and utilized by new technologies to provide data support and decision basis for exploration and development. The knowledge graph can well integrate the knowledge contained in these data and documents. However, its concept and relationship rely on manual construction, which results in its limited coverage of knowledge areas. The traditional question-and-answer (Q&A) method can get relevant answers from documents according to questions, which has the characteristics of wide knowledge coverage. Nevertheless, it is difficult to understand the contents of professional fields, which leads to its low accuracy in petroliferous basins. In order to address the above problems, this paper proposes a hybrid Q&A method for merging the knowledge graph and documents in petroliferous basins. The method takes the knowledge graph of petroliferous basins as the knowledge base of professional background. Additionally, it obtains professional-related knowledge contents from documents. In particular, to answer the question on petroliferous basins, the method firstly extracts entities from the question according to the knowledge graph. Then, with these entities, the method converts the question into a query in the knowledge graph, obtaining partial candidate answers to the question. For obtaining candidate answers from documents, the method constructs a deep semantic matching model which incorporates knowledge graph embedding. The model can match the question and answers in documents base on the information from the knowledge graph. Finally, the method leverages a sort algorithm to reorder the above two types of candidate answers from the knowledge graph and documents respectively. Compared with traditional Q&A methods, the hybrid Q&A method supports professional Q&A scenarios for the knowledge graph and documents of petroliferous basins, improving the efficiency of users' knowledge query and increasing the recall rate while ensuring the retrieval accuracy. The hybrid Q&A method has the characteristics of convenient operation, strong interaction and high accuracy, etc., which provides a platform on knowledge deep sharing and application for the study of petroliferous basins.

**Keywords:** Petroliferous Basins · Hybrid Q&A Method · Deep Semantic Matching · Knowledge Graph Embedding; Reorder

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 245–257, 2024.
https://doi.org/10.1007/978-981-97-0272-5_21

## 1 Introduction

Petroliferous basins are natural places where oil and gas occur. Many research institutions have accumulated a vast amount of knowledge and documents on global petroliferous basins [1], and there are also a large number of published literature online (such as Wikipedia). In order to further mine and utilize these knowledge documents, mature, advanced, and applicable information technologies should be utilized. Intelligent Q&A based on knowledge graphs is a series of theories and methods that organize data and mine knowledge based on graphs. It is very suitable for deep utilization of knowledge documents in petroliferous basins with complex business processes and intensive knowledge.

The knowledge of a petroliferous basin can be standardized and integrated to form a document library and knowledge graph of the basin. Due to the lack of correlation in the construction of the two, they can only be provided as a single knowledge system for users, and cannot achieve the retrieval and display of comprehensive data and knowledge resources under user query needs. Therefore, many organizations and scholars have committed to researching and improving Q&A methods. (1) Knowledge Graph Q&A (KGQA) [2, 3] refers to the process of inferring answers based on the information in a graph, which mainly includes two methods: information retrieval and semantic parsing. For example, Qu et al. [4] proposed the AR-SMCNN model to answer single relationship questions, while Luo et al. [5] proposed generating candidate query graphs to obtain answers in the knowledge graph. The knowledge graph Q&A method can clarify the reasoning process and obtain accurate answers, but its coverage of knowledge domains is limited. (2) Document-based Q&A [6, 7] refers to obtaining answers from text paragraphs based on questions, mainly including three stages: question processing, text retrieval, and answer extraction. For example, Seo et al. [8] proposed a BiDAF model for multi-stage hierarchical processing and extracting answers from documents. The document-based Q&A method can obtain answers from a wider range of knowledge, but it cannot deeply understand the contents of the professional domain, and the accuracy of answers in the professional domain is relatively low. In addition, the intelligent Q&A method has not yet been effectively implemented in the professional domain of petroliferous basins. In response to the above issues, this paper proposes a hybrid Q&A method for knowledge graph and documents of petroliferous basins, which conducts research from two aspects: the document Q&A method with embedding knowledge graph, and the comprehensive sorting method with graph and document answers, to provide users with answers that meet their needs.

## 2 Method Framework

The overall framework of the hybrid Q&A method for knowledge graph and documents of petroliferous basins is shown in Fig. 1.

An example of the hybrid Q&A method is shown in Table 1, and a detailed description of the method is provided later.

The process of this paper's method is as follows: (1) Build a full-text index and semantic index for a large number of document contents, and retrieve candidate paragraphs based on the question. The number of candidate paragraphs is generally set to

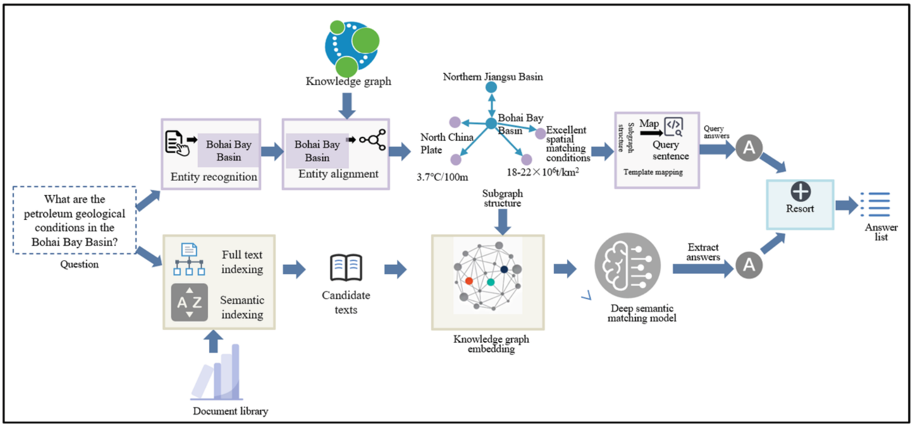


**Fig. 1.** Method framework

**Table 1.** An example of the hybrid Q&A method.

**Question: What are the petroleum geological conditions of the Bohai Bay Basin?**

**Candidate paragraphs:**

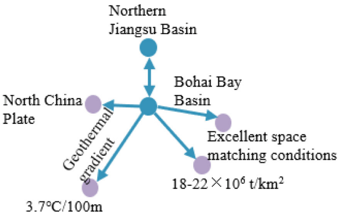
（1） ……The geothermal gradient of the Bohai Bay Basin is generally 3.7 °C/100m, and the main petroliferous sags have basically undergone a complete thermal evolution process……

（2） ……There is an order of magnitude difference in hydrocarbon generation intensity and resource abundance between the Bohai Bay Basin and the Subei Basin. The maximum hydrocarbon generation intensity in the Bohai Bay Basin is 18-22×$10^6$ t/km$^2$.

（3） ……The main source rock in the Bohai Bay Basin is also widely developed in the development period of reservoir rocks. There are at least three sets of regional caprocks in the Eocene to Pliocene. The space matching conditions of source, reservoir and caprock are excellent. The physical property sealing and pressure sealing conditions of the upper strata are far better than those of the Subei Basin……

（4） ……

**Knowledge subgraph:**



**Answers:**

(1) The geothermal gradient in the Bohai Bay Basin is generally 3.7 ℃/100m

(2) The maximum hydrocarbon generation intensity in the Bohai Bay Basin is 18-22×$10^6$ t/km$^2$

(3) Excellent spatial supporting conditions

(4) ……

5–10 manually. The knowledge graph information is fused with document information through graph embedding, and then a deep semantic matching model is used to obtain answers from the candidate paragraphs; (2) Based on the knowledge graph of petroliferous basins, analyze natural language problems and transform them into query structures that exist in the graph, and perform knowledge matching in the graph to obtain answers; (3) The answer resorting algorithm is used to measure the semantic matching degree between the question and all the above candidate answers, which is integrated into a complete and accurate answer list.

## 3 Document Q&A Method with Knowledge Graph Embedding

### 3.1 Knowledge Graph Information Embedding

**Retrieving Candidate Paragraphs.** To obtain the candidate paragraphs related to the problem from the document library accurately and quickly, this paper designs the inverted index and semantic index library. Full-text retrieval is used to obtain candidate documents based on keyword matching, while semantic retrieval is used to obtain candidate passages by local sensitivity hashing.

**Knowledge Subgraph Matching.** (1) *Domain entity recognition:* A large number of specialized concepts and knowledge are included in the domain of petroliferous basins, so the analysis of interrogative sentences requires the application of the naming entity recognition technique to identify the specialized words. This method uses the BIO annotation method to perform naming entity annotation in petroliferous basins, generates training and test sets, adds BERT [10] pre-training network as a word embedding model based on the commonly used model BiLSTM-CRF [9], and introduces attention mechanism to optimize the effect of entity recognition. (2) *Graph entity alignment:* This method uses semantic matching to establish the correspondence between the entities in the question and the graph entities. For an entity obtained from question analysis, the BERT pre-trained language model is used to convert the entity into a vector representation, and the Euclidean distance calculation between vectors and the local sensitivity hashing algorithm are used to select the aligned entities that are closest to their semantics in the graph, and a threshold is set at the end of semantic matching to filter out the semantically distant entity pairs and keep the aligned entities that satisfy the threshold.

**Knowledge Graph Embedding.** The feature vectors of candidate paragraphs and questions after embedding only contain the content representation of the context, which cannot reflect the intrinsic logical association of in-text entities with attributes and other entity relationships, and lack the semantic information at the relationship level. Since the knowledge map of petroliferous basins is a high generalization of knowledge in the text, it can be used as additional relational semantic information. The role of knowledge graph information embedding is to update the original text representation by incorporating knowledge graph information, which can obtain the relationship information between entities, and at the same time enhance the semantic representation of corresponding entities in the passage. An example of graph embedding is shown in Fig. 2.

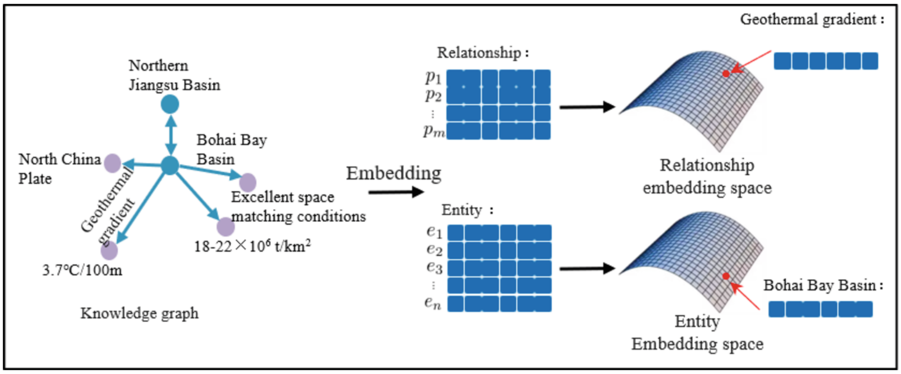


**Fig. 2.** An example of graph embedding

The fusion of the knowledge graph with the document information is shown in Fig. 3. By combining the entities “Bohai Bay Basin”, “18–22 × $10^6$ t/$km^2$”, “3.7 °C/100 m” and other expressions are integrated with the corresponding words in the paragraph expression, which enhances the semantic information of the paragraph. In which, firstly, the corresponding subgraphs in the knowledge graph of petroliferous basins are obtained by analyzing the interrogative sentences; then the vector representation of entities in the corresponding knowledge graph of the problem is obtained by a pre-trained TransE embedding model; finally, the self-attention mechanism is used to fuse the encoded text paragraph information and entity information to obtain the updated paragraph representation.

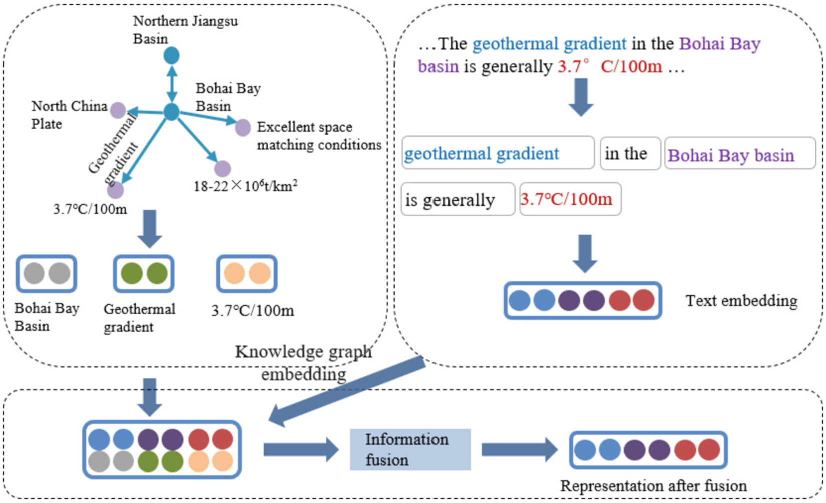


**Fig. 3.** The fusion of the knowledge graph with the document information

### 3.2 Deep Semantic Matching Model

The main goal of the deep semantic matching model is to extract the answer fragments related to the question from the candidate passages. Traditional text-based Q&A method assumes the existence of answers in candidate passages and ignores the case where the question cannot be answered, so this method overcomes this drawback by determining whether an answer exists in the passage. This method not only can accurately answer the question based on the candidate passages, but also allows for better answer identification. The structure of the deep semantic matching model is shown in Fig. 4.

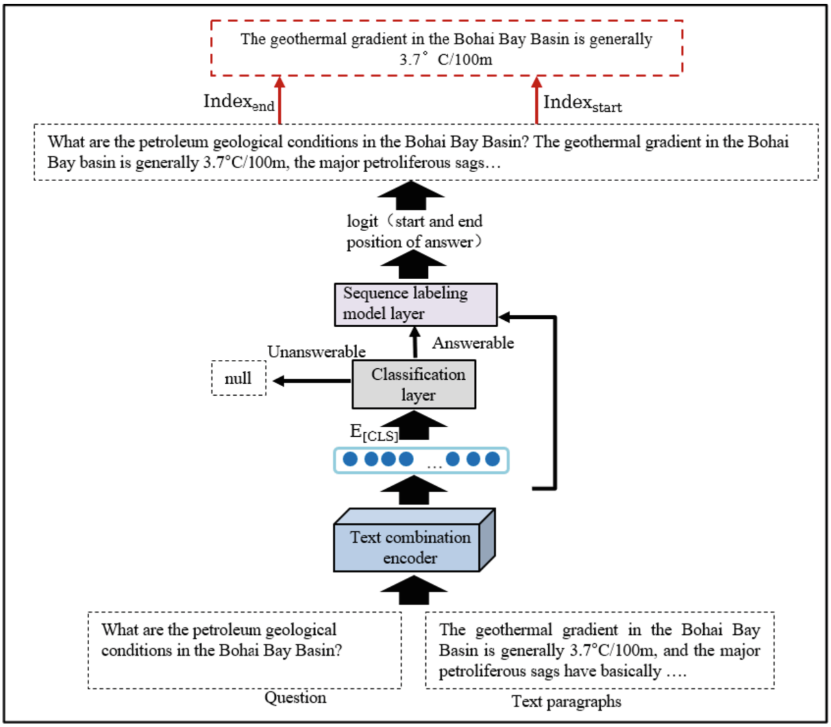

The geothermal gradient in the Bohai Bay Basin is generally 3.7° C/100m
Index$_{end}$
Index$_{start}$
What are the petroleum geological conditions in the Bohai Bay Basin? The geothermal gradient in the Bohai Bay basin is generally 3.7°C/100m, the major petroliferous sags…
logit（start and end position of answer）
Sequence labeling model layer
Unanswerable
Answerable
null
Classification layer
E$_{[CLS]}$
…
Text combination encoder
What are the petroleum geological conditions in the Bohai Bay Basin?
The geothermal gradient in the Bohai Bay Basin is generally 3.7°C/100m, and the major petroliferous sags have basically ….
Question
Text paragraphs

**Fig. 4.** Deep semantic matching model

First, the question and the candidate paragraph are jointly embedded using the BERT pre-trained language model, and the obtained encoding vector incorporates contextual information; then, the vector corresponding to the first [CLS] token of the coding sequence is represented as an aggregated sequence, and the classifier is trained to determine whether the answer to the question exists in the paragraph; for the case where the answer does not exist in the paragraph, the answer is directly set to a null value; for the case where the answer exists in the paragraph, the sequence annotation model is used for word-level prediction to determine the starting and ending position of the answer in the paragraph, and the corresponding text subsequence is extracted as the answer.

## 4 Comprehensive Sorting Method of Answers from Knowledge Graph and Document Q&A

### 4.1 Q&A Method Based on Knowledge Graph

The main goal of the Q&A method based on knowledge graph is to match answers from the knowledge graph. It is divided into three submodules: domain entity recognition, entity alignment, and template-based answer matching. The methods for the domain entity recognition submodule and the entity alignment submodule can adopt the corresponding methods in Sect. 3.1. And for the template-based answer matching submodule, the method is as follows: According to the semantic matching of entities in the question, the entity type distribution in the question can be distinguished. For the matched entities in the question, it is first judged whether the types of entities belong to the concept, instance, relation or attribute of the knowledge graph. Then, according to the types of entities in the question, the corresponding entities are filled into the manually written executable query statement. Table 2 lists the query statements corresponding to some templates.

**Table 2.** Templates for the query statements

| Types of entities | Templates | Template description |
|---|---|---|
| [instance, attribute] | match (n:INDIVIDUAL{name: $P})<br>return n.$A | Query entity attributes |
| [instance, relation] | match (n:INDIVIDUAL{name: $P})<br>match (n)-[:$R]- > (m)<br>return m.name | Query entity relations |
| [instance, instance, attribute] | match (n:INDIVIDUAL{name: $P})<br>match (m:INDIVIDUAL{name: $Q})<br>return n.$A, m.$A | Compare properties between entities |
| [instance, relation, attribute] | match (n:INDIVIDUAL{name:$P})–[:$R]- > (m) return m.$A | Query properties of adjacent entities |
| …… | …… | …… |

### 4.2 Answer Resorting

The hybrid Q&A method proposed in this paper integrates the document Q&A method based on knowledge graph embedding and the Q&A method based on the knowledge graph. By designing a resorting algorithm, the results of the two Q&A methods are integrated. The structure of the answer resorting algorithm is shown in Fig. 5.

In the research process of the answer resorting algorithm, a professional dataset is constructed for global petroliferous basins, including the set of question-answer pairs in the field. Query questions and candidate answers are embedded into a unified vector space using a language model, and the model is trained to constantly learn the semantic

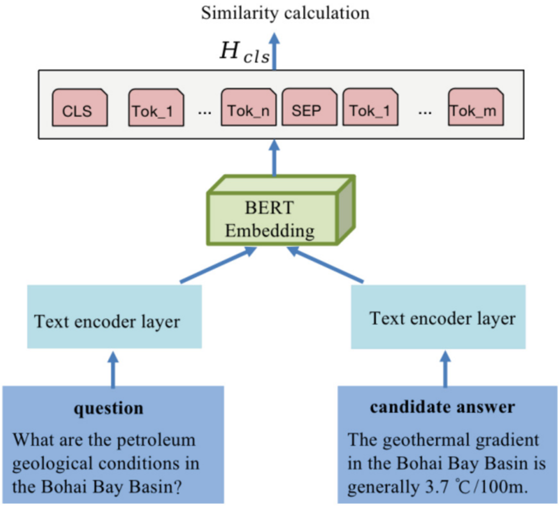

Similarity calculation
$H_{cls}$
CLS
Tok_1
...
Tok_n
SEP
Tok_1
...
Tok_m
BERT Embedding
Text encoder layer
Text encoder layer
question
What are the petroleum geological conditions in the Bohai Bay Basin?
candidate answer
The geothermal gradient in the Bohai Bay Basin is generally 3.7 ℃/100m.

**Fig. 5.** Answer resorting

matching capability of questions and answers. When the user puts forward a question, a set of candidate answers is obtained by using the above two Q&A methods, and then the semantic matching degree between the question and answer is measured according to the trained resorting algorithm. Finally, the sorted answer list is displayed according to the matching degree.

# 5 Experiments and Analysis

## 5.1 Datasets and Environment Parameters

To compare the effectiveness of the full-text retrieval and semantic retrieval, the Wikipedia dataset is used for experiments, which divides 5,035,182 training set, 10,000 verification set and 10,000 test set.

The experiment for Q&A in this paper adopts the public dataset SQuAD 11. SQuAD is an English Q&A dataset with the answer extracted from the document that includes versions SQuAD1.1 and SQuAD2.0. Among them, SQuAD1.1 contains 107,785 (question, paragraph, answer) triples, and SQuAD2.0 builds on SQuAD1.1 with more than 50,000 pieces of expanded data whose text paragraphs do not contain answers to questions.

To verify the adaptability of this method in the Chinese environment and the field of petroleum basins, this paper adopts the large-scale open dataset DuReader 12 as the basic training dataset, and customizes about 1,000 small-scale datasets for petroleum basins. By training the specific datasets through transfer learning, the Q&A method can cover petroliferous basins.

The number of the three datasets is shown in Table 3.

**Table 3.** Comparison of the number for Q&A datasets

| Set | train | dev |
|---|---|---|
| SQuAD2.0 | 130,319 | 11,873 |
| DuReader | 130,800 | 10,000 |
| The datasets for petroliferous basins | 1,000 | 100 |

### 5.2 Benchmarks and Evaluation Indicators

The term *accuracy* is used as the evaluation criterion in the hybrid Q&A method proposed in this paper. And the benchmark models for experimental comparison with the document Q&A method based on knowledge graph embedding are from QANet 13, BERT and SG-Net 14, with $EM$ 11 and $F1$ used as evaluation criteria in the experiment.

The *accuracy* is calculated by the following formula.

$$accuracy = \frac{P}{P + N} \tag{1}$$

where $P$ is the number of samples predicted correctly and $N$ is the number of samples predicted incorrectly.

$EM$ (Exact Match) is the exact match score, used to measure the exact match between the model's predicted answer and the standard answer, calculated by the following formula.

$$EM = \frac{n}{m} \tag{2}$$

where $m$ represents the total number of predicted samples, and $n$ represents the number of samples in which the predicted answer is completely consistent with the standard answer.

$F1$ is the fuzzy matching score, which is obtained by the calculated repetition between the model's predicted answer and the standard answer. The formula is as follows.

$$precision = \frac{Num_{same}}{len(predict\_answer)} \tag{3}$$

$$recall = \frac{Num_{same}}{len(gold\_answer)} \tag{4}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{5}$$

where *predict_answer* r represents the answer predicted by the model, *gold_answer* represents the actual label answer, $Num_{same}$ represents the number of repeated words between the predicted answer and the standard answer, and $F1$ represents the harmonic average of *precision* and *recall*.

### 5.3 Experimental Results and Analysis

**Analysis of experimental results of retrieval.** In this paper, two paragraph retrieval methods, full-text retrieval and semantic retrieval, are implemented for knowledge documents in petroliferous basins to narrow the range of answer candidates. The comparison results of recall are shown in Table 4.

**Table 4.** Comparison of retrieval effects

| Method | Hit@1 | Hit@10 | Hit@20 | Hit@50 |
|---|---|---|---|---|
| Semantic retrieval | 0.4772 | **0.7195** | **0.7773** | **0.8416** |
| Full-text retrieval | **0.5933** | 0.6825 | 0.6954 | 0.7090 |

The Hit@*k* in the table refers to the probability of a Hit in the first *k* results. By analyzing the results in the table, it can be seen that the recall of traditional full-text retrieval is higher in the case of single retrieval, but with the increase in the number of results, the results obtained by semantic retrieval are higher in recall and broader in coverage. Therefore, in order to consider the recall under various retrieval granularities, this paper adopts two retrieval methods at the same time to increase the recall of the whole retrieval module.

**Analysis of Resorting Experimental Results.** In order to verify the correctness of the resorting algorithm design in the hybrid Q&A method, experiments are conducted on the resorting experimental dataset based on MS MARCO 15, and the experimental results are shown in Table 5.

**Table 5.** Experimental results of the resorting algorithm

| Model | MAP | MRR |
|---|---|---|
| Initial sorting | 0.3462 | 0.3500 |
| Resorting | **0.5703** | **0.5749** |

MAP represents the average accuracy of the retrieval set, and MRR refers to the reciprocal sorting of standard answers in the search results as accuracy. Initial sorting represents the score of the basic retrieval model on the dataset. As seen from Table 5, the Resort algorithm proposed in this paper greatly improves answer ordering compared with the benchmark model, thus exerting a positive impact on the final result.

**Analysis of Experimental Results on Q&A.** *1) The Q&A method based on knowledge graph:* This method takes entity information in the KG as the answer. After the experiment on dataset of the petroliferous basins, the statistical answer accuracy reaches 82.0%.

*2) The document Q&A method based on knowledge graph embedding*: The method is trained and tested on the datasets, SQuAD2.0, DuReade, and petroliferous basins, respectively. The results of experiments on the English SQuAD2.0 are shown in Table 6. It can be seen that TextReader, the document Q&A method based on knowledge graph embedding, predicts answers on the SQuAD2.0 test set with higher exact matching scores and fuzzy matching scores than the benchmark models. For the SQuAD2.0, the baseline model assumes that the answer must exist in the text paragraph and only predicts the location of the answer within the text paragraph, not taking the absence of an answer in the text paragraph into account, so the accuracy is low. However, the method in this paper determines the answerability of the question before the answer extraction, and designs solutions for the two situations of whether there are answers in the text paragraphs, so the accuracy is higher.

**Table 6.** Experiment results on SQuAD2.0

| Model | EM | F1 |
|---|---|---|
| QANet | 73.60% | 82.70% |
| BERT | 78.01% | 80.07% |
| SG-Net | 85.10% | 87.90% |
| TextReader | **85.62%** | **88.10%** |

The method of this paper is trained on DuReader, a large-scale dataset in the Chinese open domain, to adapt to the Chinese semantic environment, and then fine-tuned and tested on the petroliferous basin dataset. In this experiment, the Chinese BERT reading comprehension model is fine-tuned on the dataset within the domain as the baseline, and the experimental results are shown in Table 7.

**Table 7.** Experimental results on the petroliferous basin dataset

| Model | EM | F1 |
|---|---|---|
| BERT | 71.25% | 80.76% |
| TextReader | **74.30%** | **82.58%** |

Table 7 shows that the document Q&A method based on knowledge graph embedding, which is called TextReader, achieves higher accuracy and recall rates on the test set of the petroliferous basin dataset compared to the benchmark model. Therefore, it is believed that this method has a certain effect on improving the accuracy of Q&A on petroliferous basins.

In addition, the experimental results also prove that in addition to the effectiveness on public datasets, this method will also have a certain degree of domain adaptability after training on specific domain datasets. Compared with the results in Table 6, it can be seen that although the Q&A performance of this method is improved compared with the baseline model, the accuracy of this method on the Chinese dataset is lower than that on the open dataset. The reason for the poor effect of the Chinese dataset is that the model cannot learn enough distribution features due to the insufficient amount of data.

*3) The hybrid Q&A method*: The hybrid Q&A method proposed in this paper combines the candidate answers obtained from the Q&A method based on the knowledge graph and the document Q&A method based on knowledge graph embedding, and finally obtains the answer list through the answer resorting algorithm. The accuracy and recall of the proposed method on Q&A in the global petroliferous basins reach 84.38% and 85.95%, which are significantly improved compared with the single knowledge graph Q&A and document Q&A, thus verifying the effectiveness of the hybrid Q&A method.

## 6 Conclusions

This study aims to design and implement a hybrid Q&A method for the knowledge graph and document library of global petroliferous basins, which mainly includes two parts: the document Q&A method with knowledge graph embedding, and the comprehensive sorting method of answers from knowledge graph and document Q&A. Through semantic analysis of the question, the corresponding subgraph is matched from the knowledge graph and the candidate node answers are inferred. Then, the graph information and the candidate paragraph information obtained from semantic retrieval are fused and inferred to obtain the candidate text answers. Finally, all the candidate answers are sorted using the resorting algorithm to generate the answer list.

This study conducts experiments on three aspects: retrieval, Q&A, and resorting. Through experimental analysis, it can be concluded that the proposed hybrid Q&A method in this paper can improve the retrieval method of candidate documents and improve the overall recall rate. It can also expand the knowledge range in petroliferous basins, while supporting two kinds of professional Q&A scenarios for the knowledge graph and documents of petroliferous basins. Compared with traditional Q&A methods, it can improve the accuracy of Q&A in global petroliferous basins. Therefore, the method researched and developed in this paper not only improves the Q&A effectiveness in global petroliferous basins, and improves the sharing level and efficiency of basin research results, but also provides researchers with better exploration and development knowledge service capabilities. In addition, as it was found in the experiments that the accuracy of this method on domain datasets is lower than that on public datasets, strengthening the adaptability of the model to professional domains is the next focus of study.

**Acknowledgments.** The project is supported by CNPC's 14th Five-Year Plan Technology Special Project (Number: 2021DJ31) and CNODC's Technology Special Project (Number: 2023-ZC-01–04).

## References

1. Mi, S., Zhang, Q., Wu, Z., et al.: Construction and application prospect of global petroleum geology and resource assessment data platform. China Pet. Explor. **2**, 38–46 (2022)

2. Liu, A., Huang, Z., Lu, H., Wang, X., Yuan, C.: BB-KBQA: BERT-based knowledge base question answering. In: Sun, M., Huang, X., Ji, H., Liu, Z., Liu, Y. (eds.) CCL 2019. LNCS (LNAI), vol. 11856, pp. 81–92. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32381-3_7
3. Li, F.-L., Chen, W., Huang, Q., Guo, Y.: AliMe KBQA: question answering over structured knowledge for e-commerce customer service. In: Zhu, X., Qin, B., Zhu, X., Liu, M., Qian, L. (eds.) CCKS 2019. CCIS, vol. 1134, pp. 136–148. Springer, Singapore (2019). https://doi.org/10.1007/978-981-15-1956-7_12
4. Qu, Y., Liu, J., Kang, L., et al.: Q&A over freebase via attentive RNN with similarity matrix-based CNN. arXiv preprint arXiv:1804.03317, 38 (2018)
5. Luo, K., Lin, F., Luo, X., et al.: Knowledge base Q&A via encoding of complex query graphs. In: Proceedings of the 2018 Conference on Empirical Methods in Nnatural Language Processing, pp. 2185–2194. ACL, Stroudsburg (2018)
6. Reddy, S., Chen, D., Manning, C.: Coqa: A conversational Q&A challenge. Trans. Assoc. Comput. Linguist. **7**, 249–266 (2019)
7. Wu, Z., Xu, H.: Improving the robustness of machine reading comprehension model with hierarchical knowledge and auxiliary unanswerability prediction. Knowl.-Based Syst. **203**, 106075 (2020)
8. Seo, M., Kembhavi, A., Farhadi, A., et al.: Bidirectional attention flow for machine comprehension. arXiv: Computation and Language (2016)
9. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)
10. Devlin, J., Chang, M.W., Lee, K., et al.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
11. Rajpurkar, P., Zhang, J., Lopyrev, K., et al.: Squad: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 2383–2392. ACL, Stroudsburg (2016)
12. He, W., Liu, K., Liu, Y., et al.: DuReader: a Chinese machine reading comprehension dataset from real-world applications. arXiv: Computation and Language (2017)
13. Yu, A., Dohan, D., Luong, M., et al.: QANet: combining local convolution with global self-attention for reading comprehension. In: International Conference on Learning Representations, pp. 1–16 (2018)
14. Zhang, Z., Wu, Y., Zhou, J., et al.: SG-Net: syntax-guided machine reading comprehension. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 9636–9643. AAAI, Menlo Park (2020)
15. Bajaj, P., Campos, D., Craswell, N., et al.: MS MARCO: a human generated machine reading comprehension dataset. arXiv: Computation and Language (2016)

# Deep Learning Study on Seismic Data Interpretation Method

Yong-hui He[1], Min Yu[1], Si-qi Ji[2], and He-ping Miao[1](✉)

[1] Shandong Yingcai University, Jinan, China
miaoheping@sdycu.edu.cn
[2] SINOPEC Key Laboratory of Geophysics, Nanjing, China

**Abstract.** With the application of deep learning algorithms in the industry, artificial intelligent technology has been developed in the field of seismic data interpretation in petroleum geophysical prospecting. This paper first starts from the analysis and research of Fully Convolutional Networks (FCN), U-Net model, the calculation of its lower accuracy results were analyzed, and the shortcomings of the model were found and pointed out; then it was proposed to introduce the High-Resolution Network (HR-Net) model into the field of intelligent interpretation of seismic data, and improve its network algorithm to make it more suitable for 3D space seismic data analysis and processing. Considering that the interpretation results of the FCN, U-Net, HR-Net algorithm cannot fully reflect the periodic phenomena and laws in the depth of the formation, the author improves HR-Net model and the high-resolution semantic fusion of the HR-Net model is also improved. The research result is the improved HR-Net algorithm model, which has certain application and promotion value in interpreting reservoirs and predicting faults from seismic image data.

**Keywords:** Deep Learning · Seismic Data Processing · FCN · U-Net · HR-Net · HR-Net Improved Model · Seismic Image Data · Reservoir Identification · Lithofacies Prediction

## 1 Introduction

With the demand of application scenarios in real life, image classification, target detection and image segmentation have become the main purpose of computer vision task processing. Among them, the classification tasks in many application scenarios are dominant in computer vision. The core of the classification task is semantic segmentation. It is a classification based on pixels. The same kind of pixels are classified into the same category. It can be said that semantic segmentation is used to process the labeled image or to understand the image. Unlike many image classification tasks, semantic segmentation is more complex and difficult. It requires a large number of pixels to support, the processing flow is more complex, and the required computing power is also multiplied. With further research of the lithology prediction, fault matching, and intelligent processing of seismic data of petroleum seismic images, the understanding or semantic

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 258–268, 2024.
https://doi.org/10.1007/978-981-97-0272-5_22

segmentation of such images is very important. It is the key technology in the field of intelligent processing of seismic images in China that needs to be solved in the field of computer vision. Therefore, it is more and more important to study how to segment the semantics and segmentation models of various seismic images efficiently and accurately.

In recent years, with the development of deep learning, it is possible to process complex and cumbersome feature extraction and result interpretation of various seismic data with artificial intelligence algorithms instead of manual work. The process is to perform convolution first, and then transpose convolution to allow the machine intelligent model to autonomously learn more detailed features of seismic images that cannot be distinguished by the human eye. Fully Convolutional Networks (FCN) was first proposed by Jonathan Long et al. in 2015 in the article Fully Convolutional Networks for Semantic Segmentation [1]. It is an algorithm framework for image semantic segmentation. The overall network model of FCN can be divided into two parts: Fully Convolutional and Deconvolution. The Fully Convolutional part borrows some classic Convolutional Neural Networks (CNN) (such as AlexNet, VGG, GoogLeNet, etc.), and replaces the final fully connected layer with convolution to extract features and form a hotspot map. The Deconvolution part is the original-size semantic segmentation image obtained by up-sampling the small-size hotspot image. However, there are many shortcomings in the resolution increased by this method. After up-sampling, semantic details will be lost, and the connection among pixels cannot be effectively calculated, resulting in semantic classification or inaccurate cutting will make it difficult to extract target features.

Based on the problems of FCN network structure, many scholars [2–9] have proposed many optimization algorithms, such as U-Net, SegNet, DeepLab-v1, etc. Among them, U-Net based on the structure of FCN proposed by Ronneberger et al. [10] in 2015 is the most remarkable. U-Net uses the construction idea of combining encoder and decoder, which is improved on the structure of FCN. The U-net algorithm is similar to the idea of FCN, and the algorithm improves the up-sampling stage. It increases the number of feature maps in the up-sampling phase (the second half of the net-work), and merges the original pixel-level low-level features into the subsequent prediction phase through jumpers, thereby increasing the accuracy of positioning. The entire network is U-shaped, with basic symmetry before and after. The feature maps on the down-sampling and up-sampling at the same level are merged by jumpers.

In 2019, Sun et al. [11] proposed HR-Net model to solve the problem of detail loss caused by up-sampling after traditional down-sampling. This network architecture realizes the repeated exchange of information and multi-scale repeated fusion by repeatedly cascading different subnets of high and low resolution, which greatly improves the utilization efficiency of spatial information and solves the problem of classification accuracy degradation caused by spatial information loss to a certain extent. The advantage of HR-Net is that it can output multi-scale feature maps, and through cross-resolution information interaction, high-resolution rich semantic feature output can be achieved. The result is that semantic details are richer and features are more obvious.

However, due to the design problem of the HR-Net model, a large amount of redundancy and several times of parameters will be generated between multiple stages [12]. In order to solve the problem of multi-scale target segmentation efficiently [13], this paper will improve and optimize HR-Net and apply it to the processing of seismic images, and further improve the processing performance by combining Object-Contextual Representation (OCR) and Atrous Spatial Pyramid Pooling (ASPP). ASPP is a method of increasing the receptive fields. ASPP can more easily obtain the semantics of the context. The dilated convolution is proposed to solve the problems in the FCN model, which improves the content aggregation and does not reduce the resolution [14]. OCR is a text recognition technology. When seismic data is processed, the model needs to identify the object features. This article is based on the principle of combining OCR and ASPP algorithms to perform image processing and obtain the feature information expression of the target context, so as to improve the accuracy of the model algorithm.

# 2 Improved HR-Net Model

## 2.1 Commonly Used Network Model Structure and Characteristics

Most of the current deep learning algorithms are processed by reducing the resolution (down-sampling), and then improving the resolution (up-sampling) to obtain feature information [15]. FCN, a fully convolutional network structure, mainly classifies images at the pixel level, but most networks are linear. Different from the classical CNN, FCN accepts the input of an image of any size, uses a deconvolution layer, and up-samples the feature map of the last convolution layer to restore it to the same size of the input image, so that a prediction can be generated for each pixel, while preserving the spatial information in the original input image, and finally, pixel-by-pixel classification is performed on the up-sampled feature map, so that each pixel can be generated [16]. As shown in Fig. 1.

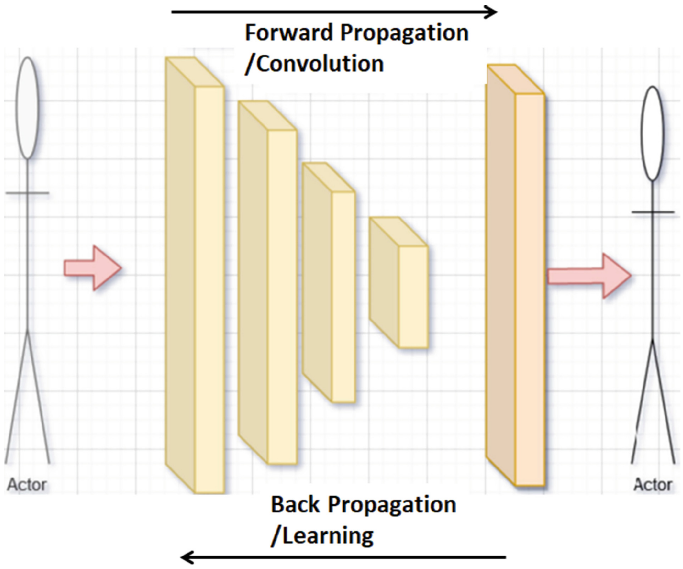


**Fig. 1.** FCN structure diagram

The characteristics of this network model are as follows:

FCN is a convolutional neural network (CNN) for image semantic segmentation. Its main structural feature is to replace the traditional fully connected layer with a fully convolutional layer, enabling the network to accept an input image of any size and output a dense prediction of the corresponding size. The characteristics of this network model are as follows:

1. Input Layer: Accepts an input image as input to the network.
2. Convolutional layer: FCN contains multiple convolutional layers for extracting features of images. These convolutional layers convolve the input image through different convolution kernels and introduce non-linearity through non-linear activation functions such as ReLU.
3. Down-sampling (pooling) layers: In order to reduce the spatial size of feature maps and preserve the main features, FCN uses down-sampling (pooling) layers. This algorithm mainly uses resolution down-sampling for semantic aggregation, enriches semantics [17], and is then used for classification tasks. Low-resolution representation learning is more friendly to classification tasks. Commonly used down-sampling methods include Max Pooling and Average Pooling.
4. Up-sampling (deconvolution) layers: In order to restore the feature map to the same size as the original image, FCN uses up-sampling (deconvolution) layers. Up-sampling can be achieved by deconvolution operations or interpolation methods.
5. Skip Connections: In order to fuse feature information at different levels, FCN introduces skip connections. Skip connections connect low-level feature maps with up-sampled high-level feature maps to provide richer contextual information [18].
6. Output layer: The last layer is the output layer, which produces a dense prediction map through convolution operations. For image semantic segmentation tasks, the output is usually a prediction map of the same size as the input image, with each pixel corresponding to a class label or class probability.

### 2.2 HR-Net Model

Compared with FCN or other network structures model, HR-Net model can achieve depth estimation and edge detection through image classification, image segmentation, target detection, face alignment, posture recognition, style transfer, image rendering, super rating and other processes[19]. HR-Net has the following characteristics:

1. Through the multi-resolution parallel flow architecture, high-resolution maintenance is achieved to ensure that spatial location information is more accurate and complete. The design concept of HR-Net continues the method of retaining a large map with resolution features and retains a large map of features. However, in the process of network advancement, it will also perform some sampling to reduce parallel feature mapping and so on. Generate multiple feature map groups with different resolutions, and then combine these feature maps to predict the segmented image and generate various functions with different resolutions. It is characterized by the fusion of the beginning, middle and last three parts of the network, not just on the final image, which can greatly improve the accuracy of the features.

2. By fusing cross-resolution information interaction features, semantic fusion is realized, and rich semantics are added on the basis of the accuracy of the original spatial information.
3. The advantage of HR-Net is that it can generate multi-scale feature maps and achieve high-resolution feature output with rich semantics through cross-resolution information interaction.
4. Keep high-resolution imaging learning: obtain feature map streams with different resolutions through parallel convolutional streams. In the main process of the whole framework, the high-resolution feature map remains unchanged, so better and more accurate location information can be obtained.
5. Multi-scale (hierarchical) fusion representative learning: Previous studies have focused on feature fusion at different levels. The whole search corresponds to the connection jump of different levels of features in the process of network function ex-traction, and realizes semantic fusion.

However, although HR-Net accurate location information is retained at this time, the semantic information in each pixel is actually limited, because the resolution compression is limited and the aggregation semantics are limited, resulting in limited performance of segmentation activities. In addition, in the previous high-resolution preservation learning, batch normalization and residual connection are not designed in the feature extraction process. According to the concept of receptive field, a pixel in the feature map is actually equivalent to multiple pixels in the original image. In other words, the semantics represented by a pixel is the semantic fusion of several original pixels. This is why the previous high-resolution preservation architecture lacks semantics. Because in order to simply maintain high resolution, the semantic information of each pixel is limited, which naturally limits the task of semantic segmentation.

Since HR-Net has no relevant literature published in seismic image processing, this paper attempts to use HR-Net algorithm model for seismic image processing for the first time. However, since the data points of seismic images are three-dimensional data, it is found that there is room for improvement in the HR-Net algorithm during model training and prediction. Therefore, this paper creatively attempts to improve the original HR-Net algorithm.

### 2.3 The Improved HR-Net (Combining ASPP and OCR)

In the HR-Net-based model, the HR-Net can be further optimized by combining with other models to increase contextual connections. This paper first assumes that there is an object context representation method, and then performs semantic segmentation. The main reason for the success of segmentation is that the label of the pixel is the label of the object where the pixel is located, and the pixel representation is enhanced by representing the corresponding object region for each pixel. Since then, this paper proposes a joint algorithm model of HR-Net + OCR + ASPP for seismic image processing, so that it can be carried out as much as possible under the training of the original HR-Net high-resolution model, and semantic segmentation is performed using the object context representation method. The ASPP module is used to capture multi-scale context information, output multi-scale contextual representations through several parallel perforated convolutions, and output cascaded representations through parallel extended convolutions. The multi-scale context scheme based on extended convolution captures multi-scale context without losing resolution. From the hypothesis of this paper, HR-Net combined with ASPP can have higher advantages and higher accuracy. The reason is that it uses pixel representation and region representation to calculate the relationship. Region representation can describe the characteristics of the target in a specific image, so the relationship between a specific image is more accurate than the simple use of pixel representation. In short, the improvement of the HR-Net original algorithm can improve the accuracy of semantic segmentation and achieve ideal results in data processing of seismic images.

## 3 Model Training and Results

### 3.1 Seismic Image Dataset

In this paper, an open source seismic data of North Sea F3 work area (as shown in Fig. 2) are used for model training, and the results are compared and analyzed on this basis.

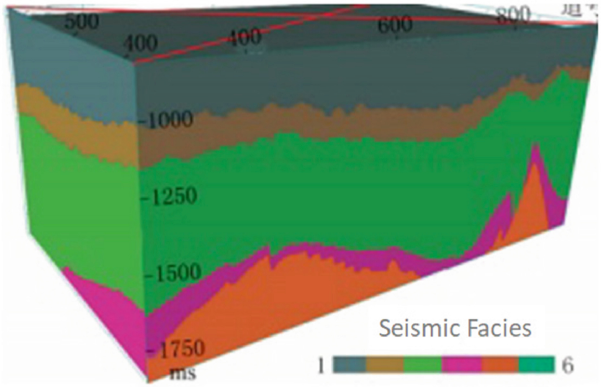


**Fig. 2.** North Sea F3 work area

### 3.2 Evaluating Indicator

For obtaining robust model, in this paper, the model is built according to the 80-10-10 ratio: 80% of the seismic image data set is used as the training set, 10% is used as the verification set, and 10% is used as the test set. This split ratio is suitable for larger datasets and simple tasks. The model has been processed and trained, and give rating indicators such as accuracy and accuracy. In the algorithm model, this paper refers to the method proposed by Yan Xing-yu et al. [20]. By introducing the mixed loss function, the objective function is added to the traditional multivariate cross entropy loss function. The Dice index loss function is used to improve the classification effect of each pixel point and the overall image segmentation effect, and overcome the problem of accuracy decline caused by uneven data distribution.

### 3.3 Model Training

The seismic image data set is divided into two blocks. The size of block 1 is 701 × 401 × 255 as the training set, and the size of block 2 is 701 × 200 × 255 as the verification set. After 10 iterations of model training, the indicators are no longer improved. This paper takes 10 epochs as the result analysis [21]. On this basis, the results of U-Net model, HR-Net model and improved HR-Net model are compared and output in the form of curve graph.

### 3.4 Experimental Design

The purpose of this paper is to compare the results of the original U-Net algorithm model with the HR-Net algorithm model and the improved HR-Net algorithm model proposed in this paper. Its seismic image F3 dataset is trained on the original HR-Net algorithm, and the results are shown in Fig. 3:

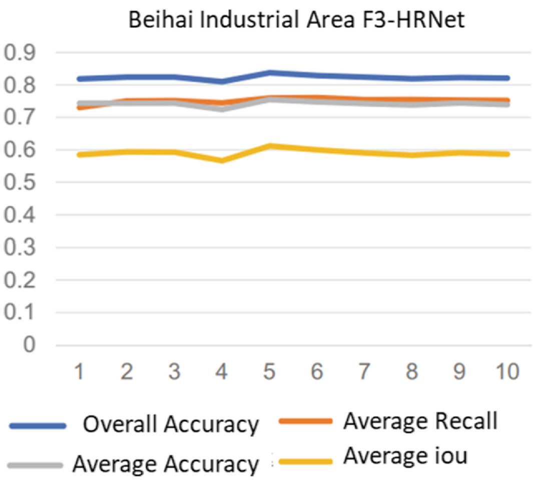


**Fig. 3.** Results of the North Sea F3-HRNet Example

The example results of the improved HR-Net model proposed in this paper are shown in Fig. 4:

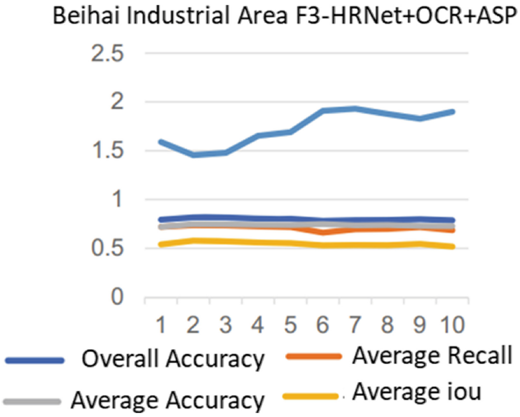


**Fig. 4.** Improved HR Net Example Results

Through comparison, it can be seen that the accuracy rate does not change much, and the highest is around 0.81, but from the seismic image, the saturation is improved, which is the sign of the improved model effect. Next, the results of the example are compared with the U-net model example. In order to reflect the experimental results more intuitively, the data of 51 × 101 × 151 provided by a work area is added here. The purpose is to identify the sand body by binary classification. Because the amount of data is small, this paper divides the last 15% as the verification set. The results of the U-net model are shown in Fig. 5.

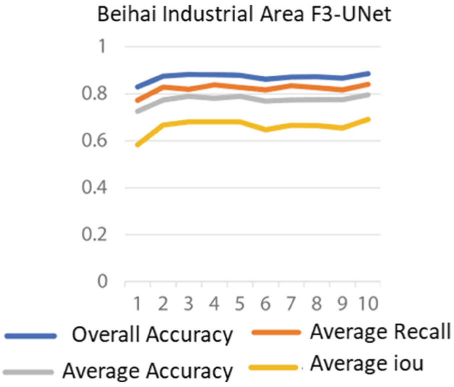


**Fig. 5.** The North Sea UNet

From the examples of the above three models, we can intuitively see the total accuracy, average recall rate and average precision rate. Through model optimization and parameter debugging, the prediction accuracy of the final model is 98% (reservoir prediction), the comparison chart of model running results as shown in Fig. 6.

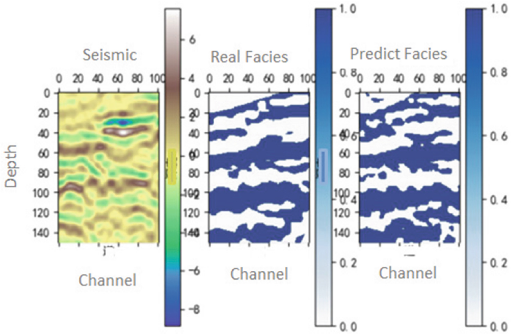


**Fig. 6.** Comparison chart of model test results

Some other useful information can also be seen on the indoor computer software model. In a word, the improved HR-Net algorithm model has formed a practical application software, which can speed up the processing of seismic data and explore a new way for the application of artificial intelligence algorithm in the field of oil exploration and development. A commercial artificial intelligence industrial application software (AiTk®) has been developed at the same time and has been applied in the market.

## 4 Conclusion

By comparing the results of FCN, U-Net, HR-Net and HR-Net improved model, this paper draws the following conclusions and suggestions:

1. U-Net model, HR-Net model and HR-Net improved model can be used to process seismic data images. In this paper, an improved HR-Net model is constructed, and its algorithm is processed by a commercial software to facilitate on-site application.
2. Compared with the original HR-Net model, the improved HR-Net model improves the processing performance.
3. It is suggested that the next research direction is to first improve the evaluation system of the model results, so that it can more fully reflect the feature attributes of the original image, and the criterion is more scientific. The second is to study the continuity of spatial points, which is the fundamental difference between seismic images and general two-dimensional images. In particular, the lateral continuity of seismic images lacks the ability to obtain global spatial prior information, and does not consider the mutual influence between pixels of seismic profiles. How to be reflected on the model is the focus of the next research, and also the key point to improve the processing of seismic data.

**Acknowledgments.** The project was supported by Sinopec Key Laboratory of Geophysics Fund Project ((Project Number: 36750000-23-FW0399-0010).

The project also was supported by Shandong Yingcai University. Fund Project Name: Lithofacies Prediction Method Based on RNN Model (Project Number: YCKY22011).

## References

1. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
2. Gao, K., Huang, L., Zheng, Y.: Fault detection on seismic structural images using a nested residual U-net. IEEE Trans. Geosci. Remote Sens. **60**, 4502215 (2022). https://doi.org/10.1109/TGRS.2021.3073840
3. Min, F., Wang, L., Pan, S., Song, G.: D2UNet: dual decoder U-net for seismic image super-resolution reconstruction. IEEE Trans. Geosci. Remote Sens. **61**, 5906913 (2023). https://doi.org/10.1109/TGRS.2023.3264459
4. Li, Z., Sun, N., Gao, H., Qin, N., Li, Z.: Adaptive subtraction based on U-net for removing seismic multiples. IEEE Trans. Geosci. Remote Sens. **59**(11), 9796–9812 (2021). https://doi.org/10.1109/TGRS.2021.3051303
5. Wang, B., Li, J., Luo, J., Wang, Y., Geng, J.: Intelligent deblending of seismic data based on U-net and transfer learning. IEEE Trans. Geosci. Remote Sens. **59**(10), 8885–8894 (2021). https://doi.org/10.1109/TGRS.2020.3048746
6. Vu, M.T., Jardani, A.: Convolutional neural networks with SegNet architecture applied to three-dimensional tomography of subsurface electrical resistivity: CNN-3D-ERT. Geophys. J. Int. **225**(2), 1319–1331 (2021)
7. Vu, M.T., Jardani, A.: Convolutional neural networks with SegNet architecture applied to three-dimensional tomography of subsurface electrical resistivity: CNN-3D-ERT. Geophys. J. Int. **225**(2), 1319–1331 (2021)
8. Vu, M.T., Jardani, A.: Convolutional neural networks with SegNet architecture applied to three-dimensional tomography of subsurface electrical resistivity: CNN-3D-ERT. Geophys. J. Int. **225**(2), 1319–1331 (2021)
9. Fu, H., Fu, B., Shi, P.: An improved segmentation method for automatic mapping of cone karst from remote sensing data based on DeepLab V3+ model. Remote Sens. **13**, 441 (2021). https://doi.org/10.3390/rs13030441
10. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
11. Sun, K., Xiao, B., Liu, D., et al.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5693–5703 (2019)
12. Dunham, M., Malcolm, A., Welford, J.: Toward a semisupervised machine learning application to seismic facies classification. In: 82nd Annual International Conference and Exhibition, EAGE, Extended Abstracts (2020). https://doi.org/10.3997/2214-4609.202011486
13. Fashagba, I., Enikanselu, P., Lanisa, A., Matthew, O.: Seismic reflection pattern and attribute analysis as a tool for defining reservoir architecture in 'SABALO' field, deep-water Niger Delta. J. Petrol. Explor. Product. Technol. **10**, 991–1008 (2020). https://doi.org/10.1007/s13202-019-00807-1
14. Qayyum, F., Betzler, C., Catuneanu, O.: The wheeler diagram, flattening theory, and time. Mar. Petrol. Geol. **86**, 1417–1430 (2017)
15. Qayyum, F., Betzler, C., Catuneanu, O. Space-time continuum in seismic stratigraphy: Principles and norms. Interpretation **6**, 1–42 (2017)
16. Kaur, H., et al.: A deep learning framework for seismic facies classification. In: First International Meeting for Applied Geoscience & Energy, SEG, Expanded Abstracts, pp. 1420–1424 (2021). https://doi.org/10.1190/segam2021-3583072.1

17. Kaur, H., Zhong, Z., Sun, A., Fomel, S.: Time-lapse seismic data inversion for estimating reservoir parameters using deep learning. Interpretation **10**(1), T167–T179 (2022). https://doi.org/10.1190/INT-2020-0205.1
18. Kaur, H., Pham, N., Fomel, S.: Separating primaries and multiples using hyperbolic Radon transform with deep learning. In: 90th Annual International Meeting, SEG, Expanded Abstracts, pp. 1496–1500 (2020). https://doi.org/10.1190/segam2020-3419762.1
19. Liu, Z., Cao, J., Lu, Y., Chen, S., Liu, J.: A seismic facies classification method based on the convolutional neural network and the probabilistic framework for eismic attributes and spatial classification. Interpretation **7**(3), SE225–SE236 (2019). https://doi.org/10.1190/INT-2018-0238.1
20. Yan, X.-y., et al.: Intelligent identification of seismic facies based on improved deep learning method. **55**(06), 1169–1177+1159 (2020). https://doi.org/10.13810/j.cnki.issn.1000-7210.2020.06.001
21. Pham, N., Fomel, S.: Uncertainty estimation using Bayesian convolutional neural network for automatic channel detection. In: 90th Annual International Meeting, SEG, Expanded Abstracts, pp. 3462–3466 (2020). https://doi.org/10.1190/segam2020-3427239.1

# Research on the Method of Inverting Indicator Diagram with Electrical Parameters of Pumping Unit Based on Neural Network

Qiao-ling Dong[1,2(✉)], Chun-long Sun[1,2], Chao Gao[3], Zhen-chao Guo[1,2], Cui Wang[1,2], Lu-fang Zhou[1,2], Xing Qi[1,2], Chun-hong Li[1,2], Hai-qun Yu[1,2], and Feng Wei[1,2]

[1] Research Institute of Oil Production Engineering, Daqing Oilfield Company Limited, Daqing, China
565817477@qq.com

[2] Heilongjiang Provincial Key Laboratory of Oil and Gas Reservoir Stimulation, Daqing, China

[3] No. 3 Production Plant of Daqing Oilfield Company Limited, Daqing, China

**Abstract.** In view of the problems that the load cell used to test the indicator diagram in the pumping unit is easy to drift in the long term and needs manual regular maintenance, a new method is proposed to demonstrate the indicator diagram of the electrical parameters in the pumping unit well. By combining neural network and big data analysis technology, BP neural network model is established to carry out learning、 training and simulation analysis on the historical data of pumping unit, to find the corresponding relationship between electrical parameters and indicator diagram, and to realize the direct conversion of indicator diagram using electrical parameters. After145 field tests, the accuracy of electrical parameter inversion diagram based on neural network reaches 93.2%. This method has the advantages of low model complexity, fast operation speed and high accuracy, which provides a new way to obtain the indicator diagram of pumping unit well, and has great significance for the digital construction of pumping unit well.

**Keywords:** Neural network · Electrical parameter · Indicator diagram · Digitization

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20–22 September 2023.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 269–279, 2024.
https://doi.org/10.1007/978-981-97-0272-5_23

## 1 Preface

The indicator diagram of the beam pumping unit is a geometric figure composed of the load and displacement of the suspension point of the rod. It is an important means to directly reflect the working condition of oil well, and can master the running state of pumping unit accurately and timely [1–3]. At present, the indicator diagram is mainly measured directly by the sensor, that is, load sensor and displacement sensor are installed at the suspension point of the pumping unit, and the load and displacement of the measured suspension point are drawn as indicator diagram. However, in the long-term working process of load sensor, due to the action of alternating load, precision misalignment or damage is easy to cause load data drift distortion, so we need manual calibration regularly, and maintenance workload is large. In view of the problems existing in the direct measurement of indicator diagram, many scholars at home and abroad use the electric power or speed of the pumping unit motor to calculate the suspension point load and obtain the indicator diagram indirectly according to the transmission and dynamic relationship of each pumping unit mechanism. Gibbs [4] proposed the indirect detection method based on the motor speed indicator graph in 1988, and calculated the load using the torque factor method according to the mechanical dynamic characteristics of the motor. Chen Pei-yi, Jiang Lin et al. [5–7] comprehensively considered the effects of transmission efficiency, inertia and balance of the pumping unit, analyzed the precise dynamics of the pumping unit, established the mathematical model corresponding to the suspended point load and electric power, and formed the power conversion method to solve the suspended point load. Zhu Zhao-kun [8] considered the influence of power loss in the modeling process, including the loss of the motor itself, and the influence of the loss of the four-link mechanism, belt and reducer in the transmission in the indicator diagram model, and established the indicator diagram fitting model. In the above indirect methods of obtaining the indicator diagram, the torque factor method tends not to converge at the upper and lower dead points, the suspension point load calculation results have a large error, the power conversion method is easy to be affected by the geometric dimension accuracy of the pumping unit, and the power loss method cannot ensure that the power loss remains unchanged when the working condition of the oil well pump changes, thus affecting the calculation accuracy of the load. It fails to meet the requirement of engineering technology for indicator diagram of pumping unit.

In recent years, because of its complex dynamic characteristics, learning, association, memory and other functions, as well as its high self-organization, adaptive ability and flexibility, artificial neural network has a wide range of practicability in automatic control, computer and artificial intelligence and other fields [9, 10], and has a broad application prospect in the field of automation, digitalization and intelligent technology of oil production system in oilfield development [11, 12]. BP neural network is the most widely used neural network at present [13]. It can summarize the internal rules of the system through the sample data provided in advance, and obtain the corresponding output results from the input of the new sample data. Due to the complex relationship between the mechanical transfer system of the pumping unit from the motor to the suspension point, such as belt, reducer, four connecting rod and so on, this paper proposes to use BP neural network to establish the electrical parameter inverse indicator diagram model, which is used to express the mechanical transfer nonlinear system of the pumping unit.

In other words, the BP neural network is studied, trained and simulated by using the electrical parameters of pumping unit well and the historical data of indicator diagram, and the mapping relationship between input and output is found to form a network training model. The power diagram is inverted and produced by BP neural network training model using the electrical parameters to be measured, which provides a new idea for indirect acquisition of indicator diagram of pumping unit.

## 2 The Principle of Research Methods

### 2.1 BP Neural Network

BP neural network is a concept put forward by scientists led by Rumelhart and McClelland in 1986 [14]. It is a multi-layer feedforward neural network trained according to the error backward propagation algorithm. It has the ability of arbitrarily complex pattern classification and excellent multidimensional function mapping ability. The basic idea of BP learning algorithm is the gradient descent method, the purpose is to obtain the minimum value of the error function, the weight of the multi-layer feedforward neural network can be adjusted in the training, so that the predicted output keeps approaching the expected output [15]. BP neural network can not only perform nonlinear approximation to any function, but also has good generalization ability.

BP network has input layer, hidden layer and output layer in structure. The network topology is shown in Fig. 1.

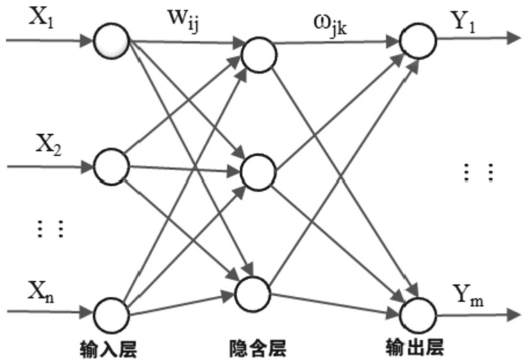


**Fig. 1.** BP neural network topology structure diagram

$X_1$, $X_2$,… Xn is the input value of BP neural network, $Y_1$,$Y_2$… Ym is the predicted value, w and ω are the network weight. BP neural network can be regarded as a nonlinear function, the network input value and the predicted value are the independent and dependent variables of the function respectively. When the number of input nodes is n and the number of output nodes is m, the BP neural network expresses the function mapping relationship from n independent variables to m dependent variables.

## 2.2 The Indicator Diagram Model Algorithm Inverted by Electrical Parameter Based on BP Neural Network

The process of using BP neural network to invert indicator diagram with electrical parameters is divided into three parts: neural network model creation, model learning and training, and model prediction. The specific algorithm steps are as follows:

1. The electrical parameters and load-displacement data of pumping units in multiple cycles are obtained through field testing. Due to the possible instability and external disturbance in the test operation process, data cleaning, screening and normalization preprocessing need to be carried out. The sample data is divided into training samples of network model and test samples.
2. Create BP neural network model, establish BP neural network topology, set the number of nodes and layers of input layer, hidden layer and output layer, configure network parameters including iteration times, learning rate and error target, and initialize the network weight threshold.
3. The data of N cycles are selected as training samples and input into the BP network model, where the electrical parameters including suspension point displacement, motor speed, motor power and motor current are taken as input vectors, and the suspension point load data is taken as output vectors. The output response of the sample data is compared with the actual value, and the connection weight of the network is adjusted when the test requirements are not met. After repeated correction, the error meets the set value.
4. The BP neural network model is used to predict the suspension point load of pumping unit, and the new data input samples are trained without changing the weight, so that the connection weight of the network can adapt to the short-term load change to ensure the prediction accuracy (Figs. 2 and 3).

## 2.3 Evaluation Index

1. MSE

   The error statistics used in the process of model training learning refers to the sum of squares of the difference between the expected output value and the actual output value of the network.

$$MSE = \frac{1}{mp}\sum\nolimits_{i=1}^{p}\sum\nolimits_{j=1}^{m}\left(y_{ij} - y'_{ij}\right)^2 \tag{1}$$

   $y_{ij}$- the expected network output value, $y'_{ij}$-the actual network output value, m- the number of output nodes, P-the number of training samples.

2. Calculation accuracy of indicator diagram

$$\mathbf{S_i} = \left[1 - \frac{|\mathbf{W_{i\,measured}} - \mathbf{W_{i\,calculated}}|}{\mathbf{W_{i\,measured}}}\right] \times 100\% \tag{2}$$

$$\mathbf{S_w} = \frac{\mathbf{S_1 + S_2 + S_3 + \ldots + S_N}}{\mathbf{N}} \tag{3}$$

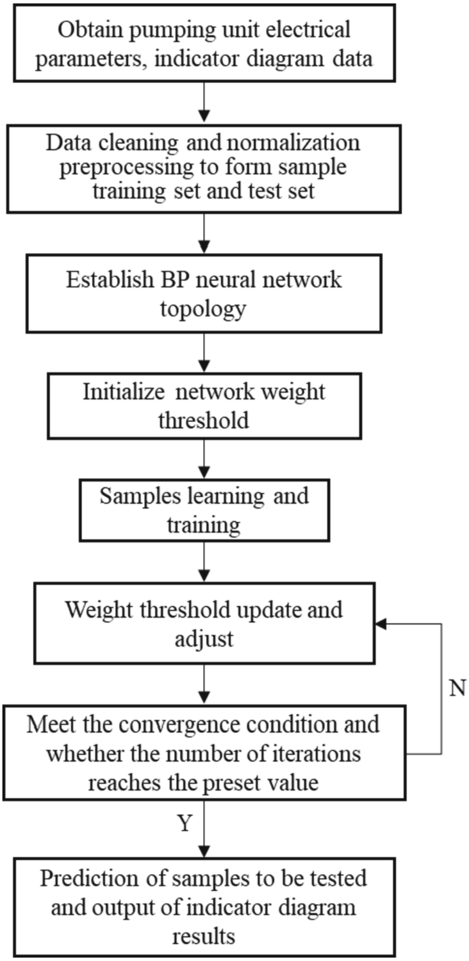

Obtain pumping unit electrical parameters, indicator diagram data
Data cleaning and normalization preprocessing to form sample training set and test set
Establish BP neural network topology
Initialize network weight threshold
Samples learning and training
Weight threshold update and adjust
N
Meet the convergence condition and whether the number of iterations reaches the preset value
Y
Prediction of samples to be tested and output of indicator diagram results

**Fig. 2.** Flow chart of the algorithm inverting indicator diagram with electrical parameter based on BP neural network

$S_i$-the single point load accuracy, %; $\mathrm{W}_{\mathrm{i\ measured}}$-the i th single point load in the actual measured indicator diagram, kN; $\mathrm{W}_{\mathrm{i\ calculated}}$-the i th single point load in the calculated indicator diagram, kN; $S_{\mathrm{w}}$-Calculation accuracy of indicator diagram, %; N-the number of points collected in a single cycle, $i = 1, 2, \ldots, N$.

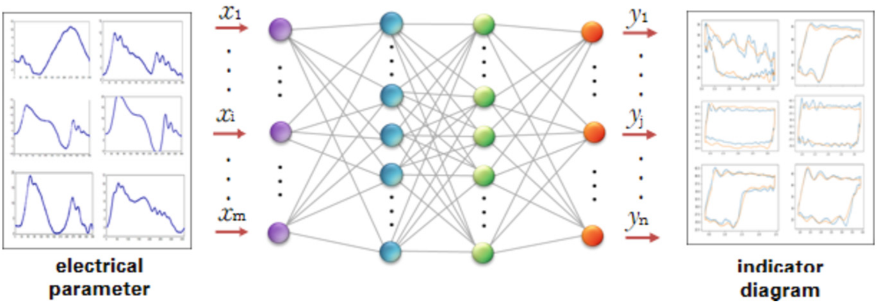


**Fig. 3.** Schematic diagram of the model inverting indicator diagram with electrical parameter based on BP neural network

## 3 Example Analysis

Based on the data source of multi-period electrical parameters and indicator diagram measured in the field of pumping unit, the model of electrical parameter inversion power graph based on BP neural network is established and data verification is carried out.

### 3.1 Data Acquisition and Preprocessing

With selecting 1000 groups of measured electrical parameters of pumping unit and corresponding data of indicator diagram, each group of electrical parameters includes motor power, motor speed and motor current, indicator diagram includes suspension point displacement and suspension point load, each parameter is periodically collected data, a cycle of 200 points, after data cleaning and screening, eliminate 30 groups of missing and abnormal data in field test. The remaining 970 groups are selected as the simulation experiment data, 825 groups with 85% data proportion are selected as the training set, and 145 groups with 15% data proportion are selected as the test set. At the same time, the data set is normalized.

### 3.2 BP Neural Network Model Establishment and Algorithm Implementation

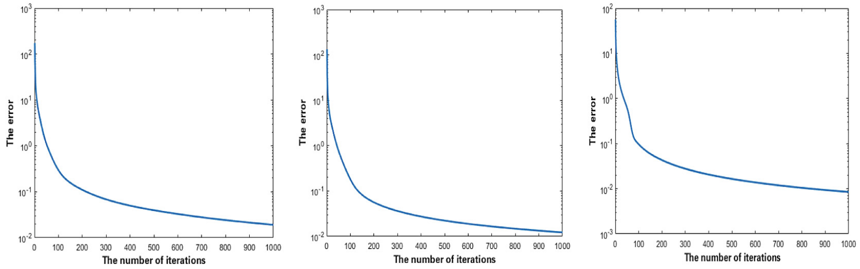
BP neural network model is created by MATLAB neural network toolbox. The specific steps are as follows:

1. The BP neural network topology is created. The suspension point displacement, motor speed, motor power and motor current are taken as input vectors, and the suspension point load data are taken as output vectors. The number of nodes in input layer and output layer are set as 4 and 1 respectively.

   It is generally believed that too many or too few neurons in the hidden layer will lead to poor performance of the neural network. The number l of nodes in the hidden layer is calculated according to the empirical formula. This example is verified by the data, when $m = 4$, $n = 1$, $a = 5$, $l = 7$, then the structure of the network is $4 \times 7 \times 1$. After calculation, $l = 10$ and $l = 15$ are taken respectively according to the performance of the network. The mean square error curve of network output values under different hidden layer nodes is shown in Fig. 4. When l = 10, the prediction error of the network is smaller and the performance is better than the other two conditions. Therefore, the number of hidden layer nodes is set as 10, so as to ensure the calculation speed of the model and the accuracy of network prediction.

$$l = \sqrt{(\boldsymbol{m} + \boldsymbol{n})} + \boldsymbol{a} \tag{4}$$

m-the number of input neurons; n-the number of output neurons; a-the constant between 0 and 10.



（a）*l*=7 （b）*l*=10 （c）*l*=15

**Fig. 4.** MSE curves of nodes in different hidden layers

2. Configuring neural network parameters. Because the selection of the network training function has a great influence on the error, by improving or selecting the efficient training learning algorithm, the diagnosis accuracy can be higher and the network error can be reduced. Under the condition that the network training objectives, training steps and network structure are set to be consistent, the variable learning rate BP algorithm, elastic BP algorithm, SCG algorithm, LM optimization method and quasi-Newton method training algorithm are respectively used to carry out the model training. The training results are shown in Table 1.It can be seen that trainlm function corresponding to LM optimization method is the fastest in network training, and the errors meet the requirements. Therefore, trainlm is configured as the training function, the number of iterations is 1000, the learning rate is 0.01, and the error target is 1e-5. At the same time, the network weight threshold is initialized.
3. Neural network model learning and training. A training set sample containing N cycles is selected and input into the established BP network model for learning training and simulation analysis. The output response of the sample data is compared with the actual value, and the connection weight of the network is adjusted when the test requirements are not met. After repeated correction, the calculation error or maximum number of iterations meet the set value.
4. Finally, the BP neural network model completed the training is verified by the samples of electrical parameters to be measured, and the suspended point load of the pumping unit is predicted. The indicator diagram is drawn by using the suspended point displacement and load, so as to realize the electrical parameter inverting indicator diagram.

**Table 1.** The results of different improved training algorithms

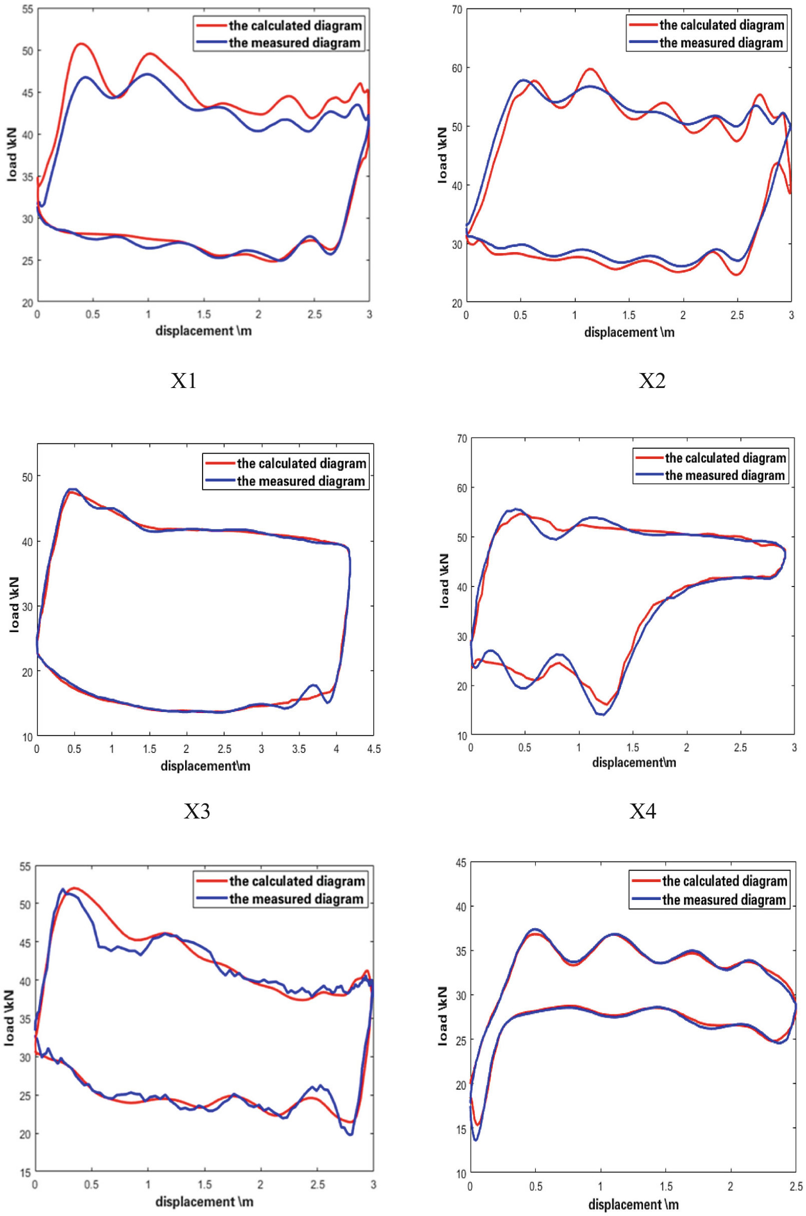
| The improved model training algorithm | The training function | The number of iterations | The training error |
|---|---|---|---|
| The variable learning rate BP algorithm | traingdx | 132 | 0.0592 |
| The elastic BP algorithm | traingdm | 1000 | 0.0793 |
| The SCG algorithm | trainscg | 121 | 0.0574 |
| The LM optimization method | trainlm | 115 | 0.0109 |
| The quasi-Newton method training algorithm | trainoss | 151 | 0.0342 |

### 3.3 Analysis of Indicator Diagram Results of Electrical Parameter Inversion

Through the verification of the established neural network model based on the electrical parameter data of the pumping unit, the predicted indicator diagram of the model is compared with the measured indicator diagram. The results of the verification example of standard well and field well (Wells X1 ~ X6) are shown in Fig. 5.

The results of the X1-X6 well electrical parameter inverting indicator diagram are compared by index according to the absolute values of minimum load and maximum load and calculation accuracy of the actual and predicted indicator diagram, as well as the average calculation accuracy of the indicator diagram, as shown in Table 2. The average accuracy of the minimum load calculation is 91.1%, and the average accuracy of the maximum load calculation is 97.5%. The average accuracy of indicator diagram calculation is 96.3%.

A total of 145 Wells are verified under different working conditions, and the average accuracy rate of indicator diagram calculation is 93.2%. It can be seen that the technology based on BP neural network has good field applicability and accuracy (Table 3).



X1

X2

X3

X4

X5

X6

**Fig. 5** Comparison between the indicator diagram calculated by the model and the measured diagram

**Table 2.** The analysis and comparison table of example results of indicator diagram inverted with electrical parameter

| Well number | Actual minimum load (kN) | Predicted minimum load (kN) | Accuracy of minimum load calculation (%) | Actual maximum load (kN) | Predicted maximum load (kN) | Accuracy of maximum load calculation (%) | Average accuracy of indicator diagram calculation (%) |
|---|---|---|---|---|---|---|---|
| X1 | 24.91 | 23.06 | 92.6 | 47.1 | 51.81 | 90.0 | 96.3 |
| X2 | 27.04 | 24.45 | 90.4 | 57.75 | 57.98 | 99.6 | 95.2 |
| X3 | 13.7 | 13.8 | 99.3 | 47.9 | 47.43 | 99.0 | 98.4 |
| X4 | 14 | 16.07 | 85.2 | 55.5 | 54.58 | 98.3 | 96.7 |
| X5 | 19.78 | 21.47 | 91.5 | 51.92 | 52.02 | 99.8 | 93.6 |
| X6 | 13.66 | 15.37 | 87.5 | 37.36 | 36.81 | 98.5 | 97.5 |
| Average | 18.85 | 19.04 | 91.1 | 49.59 | 50.11 | 97.5 | 96.3 |

**Table 3.** The accuracy result table of indicator diagram inverted with electrical parameter in different operating mode

| The pump conditions | The test number | The average accuracy of indicator diagram calculation (%) |
|---|---|---|
| Normal | 100 | 93.8 |
| Gas effect | 38 | 92.7 |
| Sucker rod breaking | 4 | 85.6 |
| Pump leakage | 3 | 90.1 |
| Total | 145 | 93.2 |

## 4 Conclusion

(1) BP neural network is used to establish the model of electric parameter inverting indicator diagram to simulate the nonlinear system of mechanical transfer in pumping unit. Through the learning training and simulation analysis of the historical data of the electrical parameters and indicator diagram of pumping unit well, the direct conversion of indicator diagram using the electrical parameters of pumping unit can be realized, which can cancel the load sensor and reduce the manual maintenance workload.

(2) The model of electrical parameter inverting indicator diagram based on BP neural network established in this paper is based on big data analysis, which needs to learn and train a large number of historical data of the pumping unit, mining the internal rules of big data. Therefore, data quality is particularly important, and data need to

be screened to eliminate the missing and abnormal data caused by the uncertainty factors in the field test, to provide a reliable data source for method validation.

## References

1. Zhang, S.R., Li, C.X.: Indirect measurement of dynamometer card of beam pumping unit. J. Huazhong Univ. Sci. Technol. **11**(32), 62–64 (2004)
2. Zhang, X., Yu, H.: Energy saving technology and development trend of pumping unit. Energy Conserv. Pet. Chem. Ind. (2), 4–6 (2007)
3. Wen, S.: Development and application of super—611 intelligent dynamometer. Jianghan Pet. Sci. Technol. **4**(2), 48–55 (2009)
4. Heck, L.P., Mcclellan, J.H.: Mechanical system monitoring using hidden Markov models. In: International Conference on Acoustics, vol. 3, no. 3, pp. 1697–1700 (2002)
5. Ma, G., Zhang, J., et al.: An introduction of cyclic detection system for pump well's working condition. Well Test. **10**(3), 68–70 (2001)
6. Chen, P.: Simulation of dynamometer card and working condition diagnosis based on pumping unit's measured electric power. Yanshan University, Qinhuangdao (2013)
7. Jiang, L.: Research on Test system of indicator diagram for Pumping unit based on electric power. Southwest petroleum university, Chengdu (2016)
8. Zhu, Z.: The research and application of indicator diagram acquisition system of beam pumping units based on the electric parameter method. Lanzhou University of technology, Lanzhou (2014)
9. Luo, X.: Algorithm and application of artificial neural network theory model, pp. 18–33. Guangxi Normal University Press, Guilin (2005)
10. Hanming, P.: The pumping unit fault diagnosis system based on the analysis of BP neural network and the indicator card. Northeastern University, Shenyang (2012)
11. Wen, B.L., Wang, Z., et al.: Diagnosis of pumping unit with combing indicator diagram with fuzzy neural networks. Comput. Syst. Appl. **25**(1), 121–125 (2016)
12. Luo, Z.: Research on soft measurement method of oil-well fluid production based on dynamometer card. Shenyang: Shenyang University of Technology (2021)
13. Yi, Y., Xiaodong, Y.A.O., et al.: Thermo-drifting error modeling of spindle based on combination of principal component analysis and BP neural network. J. Shanghai Jiaotong Univ. **47**(5), 750–753 (2013)
14. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Cogn. Model. (1988)
15. Narayanakumar, S., Raja, K.A.: BP artificial neural network model for earthquake magnitude prediction in Himalayas, India. Circuits Syst. **7**, 3456 (2016)

# Deep Learning Inversion of Electromagnetic Detection Data for Macroscopic Fractures in Croswell

Li Yin, Wei-qin Li(✉), Yan-qi Ma, and Yu-Han Wu

School of Electrical Engineering and Information, Southwest Petroleum University,
Chengdu, China
272529426@qq.com

**Abstract.** The correct identification of reservoir fractures is of great practical significance for accurately evaluating the oil and gas reserves of reservoirs and the effectiveness of hydraulic fracturing, especially for macroscopic fractures between wells with an aperture greater than 1cm, which are often the culprit of hydraulic fracturing failure. However, traditional reservoir fracture identification methods face difficulties in feature extraction and illposedness in inversion problems, making it difficult to ensure the accuracy of results. Regression prediction models based on convolutional neural networks have powerful nonlinear data mapping capabilities and can replace traditional geophysical inversion calculations. To address the issue that traditional convolutional neural networks can only handle scalar data while the actual electromagnetic field data is vector valued, this paper proposes a macroscopic fracture identification method for inter-well reservoirs based on complex-valued convolutional neural networks. By using the real and imaginary parts or amplitude and phase data of the observed field as the input to the complex-valued convolutional neural network, the information input to the network is increased, enabling the network to extract more target features and improve the identification ability of reservoir fractures. Through comparative experi-ments based on amplitude scalar data and complex-valued convolutional neural networks, the results demonstrate that the electromagnetic detection data inversion for fracture identification based on complex-valued convolutional neural networks has higher resolution and provides a new approach for the accurate identification of macroscopic fractures in inter-well reservoirs.

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20–22 September 2023.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 280–290, 2024.
https://doi.org/10.1007/978-981-97-0272-5_24

**Keyword:** MCSEM · Convolutional Neural Networks · Fractured reservoirs

## 1 Introduction

With the development of the oil and gas industry, oil and gas field exploration and development is moving towards deeper fractured reservoirs with more complex formation conditions and greater identification difficulty. Currently, nearly half of the world's explored oil and gas reserves come from fractured oil and gas reservoirs. Effective fractures can act as conduits for oil and gas transport, while ineffective fractures can impede oil and gas migration or cause hydraulic fracturing failures that impact productivity. Therefore, accurate identification of fractures is of great practical significance for predicting productivity in fractured reservoirs [1].

Hydraulic fracturing is a commonly used method to enhance reservoir permeability, which can create new fractures or enlarge existing ones within the formation [2]. As the fracturing process takes place underground and cannot be directly observed, it is necessary to diagnose and predict the positional information of the length, height, width, direction, and symmetry of the fractures after fracturing in order to determine whether the actual fractures formed are consistent with expectations. This is of great significance for the dynamic adjustment of the well network in later stage development.

The width of reservoir fractures ranges from submillimeters to tens of meters, and the distribution of fractures at different scales in underground reservoirs has an important influence on the physical properties of the reservoir and the productivity of oil and gas. The prediction of fractures is of great significance for the exploration and development of underground oil and gas resources. Macroscopic fractures refer to fractures with an opening width between centimeters and tens of meters, usually in the form of composite fractures or faults [3]. Hydraulic fracturing usually has adverse effects on macroscopic fractures between wells. As the hydraulic fracturing pressure increases, the stress on the surrounding rock formations also increases, causing damage or even fracture to the fractures. When these fractures are damaged, the stress state of the surrounding rock formations changes, which can lead to poor or even failed hydraulic fracturing results. Therefore, adequate prediction of macroscopic fractures between wells is a key factor for successful hydraulic fracturing [4].

Deep learning has excellent capabilities in nonlinear and highdimensional data mapping, and can replace traditional electromagnetic inversion processes to improve inversion accuracy and efficiency [5]. There have been many studies on using deep learning to invert large underground anomalies, but there are few studies on small and deeply buried anomalies such as reservoir fractures. Developing a deep learning electromagnetic inversion method for highresolution identification of reservoir fractures has significant practical implications for evaluating oil and gas reserves and analyzing the effectiveness of hydraulic fracturing [6].

Convolutional Neural Network (CNN) is a widely used neural network structure in deep learning. CNN uses a local connection approach, which enables it to effectively preserve spatial structure information in the input data. This is crucial when processing highdimensional electromagnetic field data, such as position, magnetic and electric field strength, and temporal frequency information [7].

The Complex-Valued Convolutional Neural Network (CV-CNN) is a type of neural network that uses complexvalued convolution kernels and feature maps, allowing for more accurate learning of signals in the complex domain. Compared to realvalued CNN, CV-CNN have certain advantages [8]. Complex numbers are a mathematical concept consisting of a real part and an imaginary part, and can be used to represent the phase and amplitude information of certain signals. Real-valued CNN use real-valued convolution kernels and feature maps to process scalar data, which can cause the loss of key information such as phase and amplitude when learning complex-valued data. Electromagnetic field data usually consists of complex-valued data with complex phase and amplitude information, and in the analysis of electromagnetic data, CV-CNN are more effective in processing complex-valued data than CNN.

## 2 Research on Forward Theory

Forward modeling is an important technique in the interpretation of electromagnetic data for geoelectric exploration. It involves solving the theoretical field values by given initial models and boundary conditions.Currently,numerical methods based geoelectric forward modeling techniques have been developed and matured, including several methods such as integral equation method, finite element method, and finite difference method [9]. Integral equation method is mainly used to solve simple models. Although it involves complex mathematical calculations, its advantage is that it only needs to solve the unknown field in the anomalous area, without partitioning the entire area, and without involving complex differential equation boundary problems [10].

The Maxwell equations are the fundamental physical laws that describe the relationships between electromagnetic fields in space. They reveal that the interaction between electric and magnetic fields is the fundamental reason for the existence of electromagnetic fields [11]. The Lippmann Schwinger equation and the scattering field distribution equation are basic equations that describe scattering phenomena, but they are in integral form and it is difficult to obtain analytical solutions for the scattering fields through direct solving. To address this issue, a forward problem calculation method based on the method of moments (MoM) is introduced, which can discretize the integral equation into a linear equation system and solve for an approximate solution [12].

Dividing the computation domain Ω into M equally sized and shaped grid cells, it can be approximated that the electric field within each grid cell is uniform when the edge length of each grid cell is less than onetenth of the wavelength. Based on this, the integral formula can be rewritten into a discrete form, thereby transforming the solving process into a problem of solving a linear equation system.

$$\overrightarrow{E}^{tot}_{p;m} = \overrightarrow{E}^{inc}_{p;m} + k_0^2 \sum\nolimits_{m'=1}^{M} A_{m'} \overrightarrow{G}(\overrightarrow{\boldsymbol{r}}_m, \overrightarrow{\boldsymbol{r}}_{m'}) \overrightarrow{J}_{p;m'} \tag{2.1}$$

where $\overrightarrow{E}^{tot}_{p;m}$ represents the total field in the m-th grid cell at the p-th incidence; $\overrightarrow{E}^{inc}_{p;m}$ represents the incident field in the m-th grid cell at the p-th incidence; $A_{m'}$ represents the size of the m-th grid cell; $\overrightarrow{G}\left(\overrightarrow{r}_m, \overrightarrow{r}_{m'}\right)$ is the free space Green's function, which satisfies:

$$\overrightarrow{G}(\overrightarrow{r}_m, \overrightarrow{r}_{m\prime}) = \frac{i}{4}H_0^1(k_0|\overrightarrow{r}_m - \overrightarrow{r}_{m\prime}|) \tag{2.2}$$

$H_0^1\left(k_0|\overrightarrow{r}_m - \overrightarrow{r}_{m'}|\right)$ is the zeroth-order Hankel function of the first kind, $\overrightarrow{J}_{p;m'}$ represents the equivalent current within the $m^{'}$ th grid cell at the p-th incidence, which satisfies the following equation:

$$\overrightarrow{J}_{p;m\prime} = (\varepsilon_{r;m\prime} - 1)\overrightarrow{E}_{p;m\prime}^{tot} \tag{2.3}$$

where $\varepsilon_{r;m\prime}$ represents the relative permittivity of the medium within the $m\prime$ th grid cell. Finally, the discretized formula is:

$$\overrightarrow{E}_p = \overrightarrow{E}_p^{inc} + \overrightarrow{G}_D\overrightarrow{\chi}\overrightarrow{E}_p \tag{2.4}$$

$\overrightarrow{G}_D$ represents the relationship between the grid cells within the computation domain $\Omega$, and satisfies:

$$\overrightarrow{G}_D = k_0^2 A_{m\prime}\overrightarrow{G}(\overrightarrow{r}_m, \overrightarrow{r}_{m\prime}) \tag{2.5}$$

$\overrightarrow{\chi}$ represents the contrast of the grid cell, and satisfies:

$$\overrightarrow{\chi} = (\overrightarrow{\varepsilon}_r - 1) \tag{2.6}$$

Therefore, Eq. (2.3) can be rewritten as:

$$\overrightarrow{J}_p = \overrightarrow{\chi}\overrightarrow{E}_p^{tot} \tag{2.7}$$

Substituting Eq. (2.4) into (2–7), we have:

$$\overrightarrow{J}_p = \overrightarrow{\chi}(\overrightarrow{E}_p^{inc} + \overrightarrow{G}_D\overrightarrow{J}_p) \tag{2.8}$$

Similarly, the integral equation for the scattered field can be discretized as:

$$\overrightarrow{E}_p^{sca} = \overrightarrow{G}_S\overrightarrow{J}_p \tag{2.9}$$

where, $\overrightarrow{G}_S$ represents the Green's formula for the relationship between the grid cells located within the computation domain $\Omega$ and the receiving antenna.

The electromagnetic scattering problem aims to obtain the scattered field data. In Eq. (2.9), $\overrightarrow{G}_S$ is a known quantity and $\overrightarrow{J}_p$ is the quantity to be solved. Therefore, the key to the electromagnetic scattering problem is to obtain the induced current data inside the mesh. By transforming Eq. (2.8) and separating $\overrightarrow{J}_p$, we can obtain:

$$\overrightarrow{J}_p = (\overrightarrow{I} - \overrightarrow{\chi}\overrightarrow{G}_D)^{-1}(\overrightarrow{\chi}\overrightarrow{E}_p^{inc}) \tag{2.10}$$

After using matrix inversion to obtain $\overrightarrow{J}_p$, we can substitute it back into Eq. (2.9) to obtain the scattered field data.

The above is the complete process of solving the electromagnetic scattering problem using MoM. Through this method, it is more convenient to solve the scattered field distribution equation and obtain the approximate solution of the scattered field. This not only improves the computational efficiency but also provides more accurate numerical simulations for the study of scattering phenomena.

# 3 Model Studies

## 3.1 Forward Model Setup

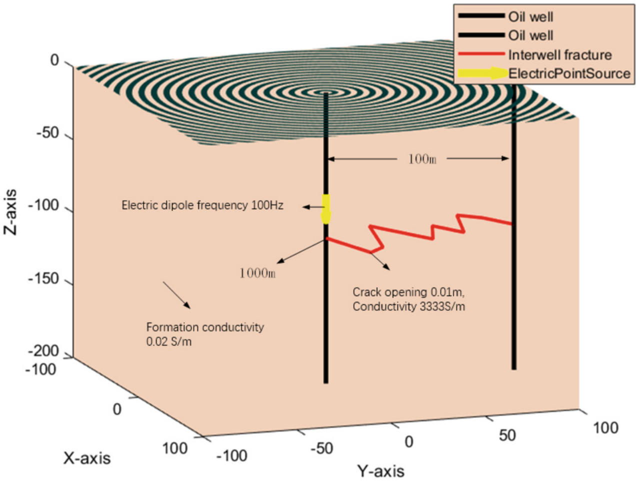


**Fig. 1.** Schematic Diagram of Forward Model

The well to well electromagnetic detection system model is shown in Fig. 1. The red curve in the figure represents the interwell reservoir fracture, with a burial depth of 1000 m, an aperture of 0.01 m (fracture diameter), and a conductivity of 3333 S/m. The yellow downward arrow represents the electric dipole transmitter with a working frequency of 100 Hz. The black line segment that contains the electric dipole represents the oil-producing well, which in this experiment represents the signal source transmitting well. The other black line segment connected to the fracture also represents an oil-producing well, and the distance between it and the transmitting well is 100 m. The diffusive circular area on the ground in black represents the receiving area of the electromagnetic field at the ground (Z = 0 m), with a reception range of 100 m to 100 m in the X direction and 100 m to 100 m in the Y direction, forming a rectangular area with a spacing of 2 m between adjacent receivers. The brown area below the black area on the ground represents the background medium of the stratum, with a conductivity of 0.02 S/m. The specific parameter settings for the forward model are shown in Table 1.

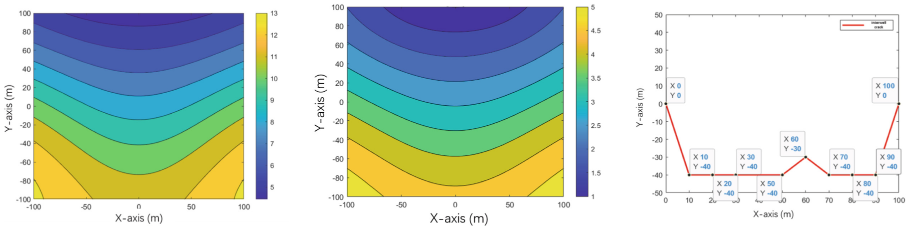
**Table 1.** Configuration Parameters of the Detection System

| Parameters | Numerical values |
|---|---|
| Transmitter type | Dipole |
| Observed component (A/m) | $H_x$, $H_y$ |
| Observation plane (m) | X = −100:2:100, Y = −100:2:100, Z = 0 |
| Transmitter frequency (Hz) | 100 |
| Fracture area (m) | X = 0:100, Y = −40:40, Z = −1000 |
| Background conductivity (S/m) | 0.02 |
| Fracture conductivity (S/m) | 3333 |

### 3.2 Synthesis of Simulation Data

A dataset of ground-based electromagnetic data and underground fracture locations was collected through electromagnetic forward simulation software. The forward model can simulate electromagnetic and fracture data under different physical parameters and experimental conditions, thereby obtaining a diversified neural network dataset, laying the data foundation for the experiment.

To ensure that the sample dataset contains rich fracture location information, the method of controlling variables was used to create the dataset, only changing the location of the fracture in each sample, while keeping other parameters constant. In the forward modeling process, a receiver array of 101*101 was used to cover the entire receiving area, and the fracture position was set to pass between two wells with a length of 100 m in the X direction. To reduce the output data size of the CV_CNN, the fracture was divided into 10 sections in the X direction, and the Y coordinate represented the position of the fracture in the opposite direction. Therefore, the output data size of each sample is 9*1. The dataset contains a total of 1000 samples, 900 of which are used as training data and 100 as test data (Fig. 2).



(a) Real part of magnetic field

(b) Magnetic field imaginary part

(c) Outputting Fracture Positions

**Fig. 2.** Dataset samples

### 3.3 Network Structure

To compare the performance of CNN and CV-CNN, it is necessary to ensure that factors such as network structure, dataset, and training parameters are the same. Only under these identical conditions can we better compare the performance of the two models and obtain more accurate conclusions. The framework of the neural network is shown in Fig. 3. The input layer C1 inputs a 101*101*2 input matrix into the network, undergoes convolution with 8 convolution kernels of size 2*2 and stride 1, and then undergoes batch normalization (BN) and rectified linear unit (ReLU) operations for the first step of feature extraction. Then, a 22 pooling operation with stride 2 is used to connect to the first layer C2 (50*50*8) of the encoding part. The same operation is repeated with 32, 64, 64, and 128 convolution kernels of size 55 and stride 1 to obtain the C3 layer (25*25*32), C4 layer (12*12*64), C5 layer (6*6*4), and C6 layer (3*3*128). After the last layer C6 of the encoder, a fully connected layer C7 with 10,000 neurons is used for high-dimensional feature mapping. Finally, an output layer C8 with 9 neurons is used to output the mapped feature vector as a matrix describing the crack position. The specific information parameters of each layer of the network are shown in Table 2. The advantage of CNN is that it can discover features in images through convolutional kernels, making it easier for classification or regression. In the CV_CNN network used in the experiment, each layer has its unique function. For example, the first layer C1 can extract the basic features of the input data, and the second layer C2 can further extract important features from the data. The subsequent layers C3, C4, C5, and C6 then gradually extract more abstract and advanced features. The fully connected layer C7 can map the features, and the output layer C8 is a matrix describing the crack position. Through these operations, the model's accuracy and generalization are improved by retaining the essential features in the data while removing some noise and redundant information.

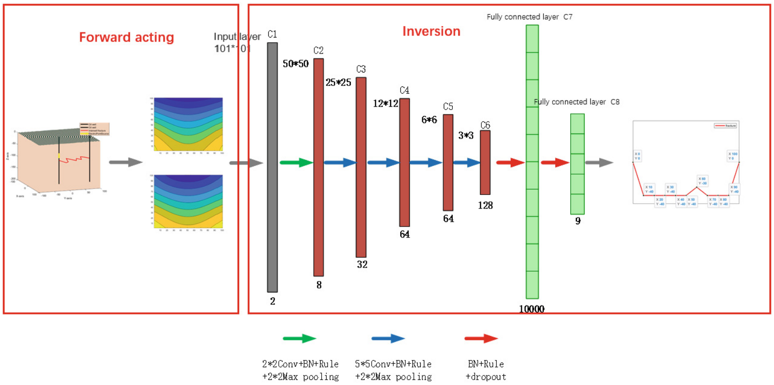


**Fig. 3.** Network architecture.

**Table 2.** Parameters information of neural network layers.

| Layer | Type | Input Size | Kernel Size | Stride | Output Size |
|---|---|---|---|---|---|
| Input | (input layer) | (101,101,2) | | | (101,101,2) |
| Conv_1 | conv | (101,101,2) | (2,2) | 1 | (101,101,8) |
| Maxpool_1 | pool | (101,101,8) | (2,2) | 2 | (50,50,8) |
| Conv_2 | conv | (50,50,8) | (5,5) | 1 | (50,50,32) |
| Maxpool_2 | pool | (50,50,32) | (2,2) | 2 | (25,25,32) |
| Conv_3 | conv | (25,25,32) | (5,5) | 1 | (25,25,64) |
| Maxpool_3 | pool | (25,25,64) | (2,2) | 2 | (12,12,64) |
| Conv_4 | conv | (12,12,64) | (5,5) | 1 | (12,12,64) |
| Maxpool_4 | pool | (12,12,64) | (2,2) | 2 | (6,6,64) |
| Conv_5 | conv | (6,6,64) | (5,5) | 1 | (6,6,128) |
| Maxpool_5 | pool | (6,6,128) | (2,2) | 2 | (3,3,128) |
| Fully Connected layer | Dense | (3,3,128) | | | 10000 |
| Fully Connected layer | Dense | 10000 | | | 9 |

The network parameters are set as shown in Table 3:

**Table 3.** Network training parameters.

| Neural network parameter | Setting value |
|---|---|
| Batchsize | 5 |
| LearningRate | $1e^{-3}$ |
| Optimizer | Adam |
| LossFunction | RMSE |
| MaxEpochs | 100 |
| LearnRateDropFactor | 0.1 |
| LearnRateDropPeriod | 100 |
| Dropout | 0.2 |

### 3.4 Inversion Evaluation

The network was trained using 1000 samples, with 90% of the samples used for network training and 10% used for network validation. The loss function value curves for both CNN and CV-CNN are shown in Fig. 4, respectively.

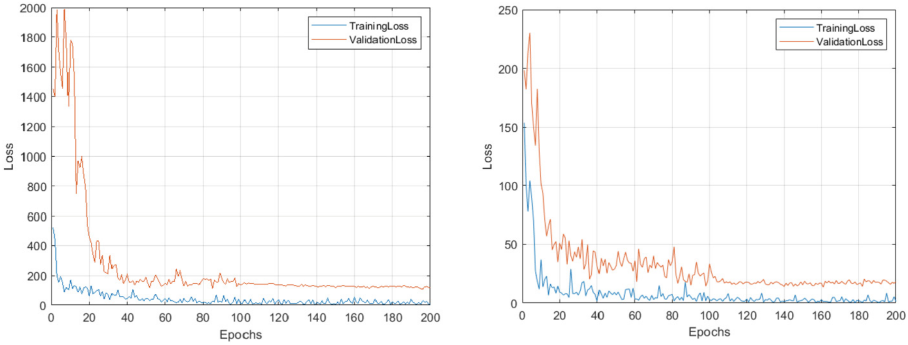


**Fig. 4.** Loss curve of CNN network training (left). Loss curve of CV-CNN network training (right).

The visualization comparison between the true labels and the predicted results of new samples by the trained networks is shown in Fig. 5. It can be seen from the figure that the CV-CNN and CNN models trained on the dataset composed of ground magnetic data and fracture information can accurately predict the location of reservoir fractures with a well spacing of 100 m, a burial depth of 1000 m, and an aperture of 0.01 m.

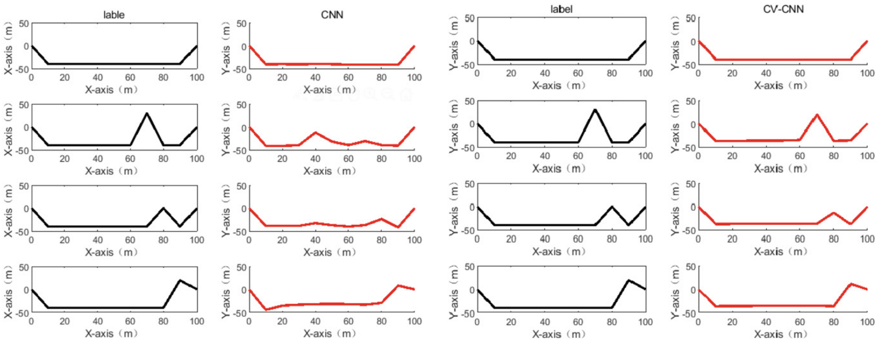


**Fig. 5.** Comparison between tags and CNN predictions (left), comparison between tags and CV-CNN predictions (right).

The results show that the deep learning model trained using the ground magnetic field data and the matrix representing the position of fractures can quickly and accurately predict the location of fractures in the 1000 m deep and 0.01 m aperture underground wells. Both the CNN and CV-CNN models can effectively learn the input data and predict the location of fractures. By comparing the loss values of the CNN and CV-CNN, it can be concluded that the CV-CNN has a better prediction result for the input electromagnetic data, and has a higher accuracy in fitting the original labels.

## 4 Conclusion

We propose a new method based on combining the CV-CNN with electromagnetic detection technology to predict interwell fractures, addressing the difficulties in feature extraction and ill-posed inverse problems associated with traditional reservoir fracture identification methods. This method is more efficient and accurate than traditional exploration methods, and can better explore the distribution of underground inter-well fractures, avoiding the problem of hydraulic fracturing failure due to inaccurate identification of macroscopic fractures.

Firstly, a forward inter-well reservoir fracture model was designed to meet actual engineering conditions. Inter-well fractures with a depth of 1000 m, an aperture of 0.01 m and a length of 100 m were set up, and an inter-well electromagnetic detection system with well to ground transmission and reception was arranged. The emission frequency of the electric dipole was set to 100 Hz, taking into account the transmission characteristics of the electromagnetic field in the reservoir. Then, CV-CNN and CNN neural networks were built to learn from the scattered field data of inter-well fractures obtained through forward simulation. The experimental results show that the CV-CNN model has better learning ability for scattered field data of fractures than the CNN model. The CV-CNN model can quickly and accurately predict the location of inter-well fractures with a depth of 1000 m and an aperture of 0.01 m from the ground magnetic field data. The application of this model has broad prospects in oil and gas reservoir evaluation, hydraulic fracturing monitoring, geological exploration, underground water resources management and other fields. In addition, this study also provides a reference for using deep learning technology to carry out underground fracture detection research.

**Acknowledgments.** The project is supported by Sichuan Provincial Science and Technology Plan, 2020YFG0182, Downhole wireless low-frequency electromagnetic signal high-speed two-way transmission system in oil casing environment; National Natural Science Foundation of China, 41974117, Research on the Micro-nano-pore dense rock conduction-dielectric mechanism and full-spectrum detection theory.

## References

1. Teng, Y.: Study on response characteristics of crack in electromagnetic remote detection. Xi'an Shiyou University (2021). https://doi.org/10.27400/d.cnki.gxasc.2021.000846
2. Zeng, L., Su, H., et al.: Fractured tight sandstone oil and gas reservoirs: a new play type in the Dongpu depression, Bohai Bay Basin, China. AAPG Bull. **97**(3), 363–377 (2013)
3. He, Z., Hu, G., Huang, D.: Seismic identification and corresponding strategy of fractured development zone in tight reservoir. Oil Geophys. Prospect. (02), 190–195+122–252 (2005)
4. Jia, J., Wang, D., Li, B.: Study on influencing factors of effective fracturing radius in hydraulic fracturing. J. Saf. Sci. Technol. China **18**(06), 58–64 (2022)
5. Tong, Z., Gao, J., Yuan, D.: Advances of deep learning applications in ground-penetrating radar: a survey. Constr. Build. Mater. **258**, 120371 (2020)
6. Wang, J.: Lecture on nonlinear inversion methods for geophysical data (II) Monte Carlo method. J. Eng. Geophys. **02**, 81–85 (2007)

7. Wang, H.: Study on large-scale gravity magnetic data inversion based on parallel computing and deep learning algorithm. Jilin University (2020). https://doi.org/10.27162/d.cnki.gjlin.2020.004090
8. Ma, Z.: Study on electromagnetic scattering and inverse scattering based on deep learning. Hangzhou Dianzi University (2021). https://doi.org/10.27075/d.cnki.ghzdc.2021.000287
9. Wang, H.: Study on geoelectric electromagnetic forward modeling based on physical information neural network. Jilin University (2022). https://doi.org/10.27162/d.cnki.gjlin.2022.001262
10. Tang, J., et al.: Forward modeling of controllable source electromagnetic integral equation for complex underground abnormal bodies. J. Geophys. **61**(04), 1549–1562 (2018)
11. Li, Z.: 3D Geoelectric electromagnetic forward modeling under undulating terrain conditions. Guilin University of Technology (2022). https://doi.org/10.27050/d.cnki.gglgc.2022.000224
12. Mengnan, Z.: Study on Electromagnetic Scattering from Rough Surfaces Based on Single Integral Equation. Anhui University (2021). https://doi.org/10.26917/d.cnki.ganhu.2021.000187

# Comprehensive Management and Application for Big Data of Pre-stack Seismic Data

Hong-wei Deng(✉), Hai-bo Zhao, Zhi-ming Zhang, Sheng Ding, Hai-hong Chu, and Bin Chen

Exploration and Development Research Institute of Daqing Oilfield Company Ltd., CNPC, Daqing 163712, China
dhw2013@petrochina.com.cn

**Abstract.** As the revitalization and development of Daqing oilfield has entered a new stage, new requirements have been put forward for the large-scale processing of massive seismic data, and the seismic data processing has entered the era of basin-level massive pre-stack data. The quality of massive pre-stack seismic data is closely related to the quality of migration imaging, which greatly affects the optimization of trap target, well position deployment and adjustment of engineering scheme. Therefore, the comprehensive treatment of massive prestack seismic data is of great significance to the high-quality development of oilfield exploration and development. Due to issues such as missing data header information, non-standard archive content, difficulty in data back-tracking, unsynchronized data structure and document storage, the quality of pre-stack seismic data is affected to varying degrees, unable to meet the needs of high-quality, efficient, and full lifecycle exploration and development. Professional big-data comprehensive management means are needed. Through continuous exploration and practice, a "Five Refined Quality Control" has been established for pre-stack big-data management methods, which includes institutionalized modes, quality control processes, evaluation forms, software standardization, and management informatization, achieving standardized management and standardized governance of pre-stack seismic data. The method is applied to massive seismic pre-stack data, and established a pre-stack middle results database for the entire exploration area of Daqing. In the continuous processing task of tract 24 in the Gulong area of the Songliao Basin, high-quality

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20–22 September 2023.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 291–297, 2024.
https://doi.org/10.1007/978-981-97-0272-5_25

middle results data were used to shorten the 26 month continuous processing cycle to 9 months, greatly improving the seismic processing efficiency and achieving significant promotion and application results.

**Keywords:** Seismic data processing · Seismic pre-stack big-data · Full lifecycle · Middle results database · Processing efficiency

## 1 Introduction

The revitalization and development of Daqing Oilfield during the 14th Five Year Plan period has entered a new stage [1]. Faced with the high-quality development goals of the oilfield, whether it is from early basin evaluation to precise exploration of oil and gas rich depressions, or from reservoir evaluation deployment to development plan formulation and residual oil potential tapping in the later stage of development, the exploration targets are becoming increasingly complex and the exploration difficulty is also increasing. In Songliao conventional oil field, the remaining oil reservoirs are small in scale, complex in structure, and strong in concealment. It is necessary to strengthen the large-scale recognition of geological structure. In the field of Songliao shale oil [2], the enrichment layer is thin, the horizontal heterogeneity is strong, and the seismic reflection characteristics have significant differences, resulting in insufficient understanding of large-scale geology. It is difficult to identify gas-bearing sandstone and small faults in the natural gas field in Sichuan and Chongqing, with large depth of marine strata, thin dolomite reservoirs, large depth of fractured reservoirs and complex wave field propagation processes. It is difficult to identify thin interbedded lithologic reservoirs and lithologic structural composite reservoirs in the field of tight oil in Songliao. In the field of deep gas in Songliao, the deep fracture structure, source rock distribution and volcanic rock distribution should be recognized. Tight oil has strong tectonic activity and complex geological conditions in the Hailar and surrounding areas, and the remaining exploration targets are “fragmented, low, scattered, and deep”. Different exploration targets and complex geological needs urgently require seismic processing results at the basin level, zone level, and trap level. Seismic data processing has entered the era of massive prestack in basin level [3].

The quality of big data in pre-stack seismic data is closely related to the quality of migration imaging, and greatly affects the timeliness of trap target selection, well location deployment, and engineering scheme adjustment. Before drilling, seismic data should be applied to timely, accurate, and high-precision support well location verification and deployment. During drilling, seismic data should provide real-time adjustment basis for point-to-point and complex targets based on complex geological conditions encountered during the drilling process. After drilling, seismic data is also applied to predict and optimize engineering construction parameters. However, due to limited space storage in the past, the pre-stack data of seismic processing has not been archived and stored, which has greatly affected the quality and applicability of the data, resulting in waste of data resources and a large amount of repetitive processing work. Therefore, the comprehensive management of pre-stack seismic big-data and the standardized management of seismic intermediate results data is a practical and feasible technical approach to achieve deep exploration data and improve production efficiency. It is of great significance for the high-quality development of oilfield exploration and development.

## 2 Analysis

With the continuous development of geophysical acquisition technology, 3D seismic data of high density and fold, such as "two wide and one high", vibrator, and single-point high-precision detectors, are constantly emerging, so the amount of seismic data is assuming an exponential growth. After analysis and induction, it has been found that existing pre-stack seismic data often suffers from issues such as missing trace header information, non-standard archive content, uncontrolled data backtracking, and asynchronous data structure and document storage. The problem can be summarized as follows:

(1) The data information is incomplete, especially the basic information of the project, and there are missing items in the collection and processing parameters, technical indicators, workload, assessment indicators, etc.
(2) The archived materials did not follow the corresponding standards, lacked the required archived results, and the project results storage rate was incomplete.
(3) The data quality is not high, and there are various types of errors, such as inaccurate point positions, duplicate lines, outliers in the data, inconsistency between file names and content, and garbled data.
(4) The correlation between seismic acquisition, processing, and interpretation has not been established, data backtracking cannot be conducted.
(5) The structured storage of key data is not synchronized with document storage.
(6) Online and offline archiving data are inconsistent and synchronized.

## 3 Method

Guided by the "Revitalization and Development" outline of the oilfield, in response to the massive amount of seismic big-data, a unified seismic data format was adopted, a complete basin level pre-stack large database was constructed, meanwhile the "Five modernizations of fine quality control" method was created to achieve comprehensive management of pre-stack seismic data with high efficient and speed. The specific measures are as follows:

(1) Adopting a cross-multi-task quality control mode to institutionalize quality control work. Based on the seismic processing process, the "cross-multi-task quality control" mode is formulated to strengthen the quality control of intermediate processing, to implement a processing quality control processing work system, and to ensure the efficient operation of quality control tasks. During the quality control process, one person controls multiple processing processes to ensure a reasonable allocation of workload. At the same time, adjacent processes in the same work area cannot be controlled, reducing the subjective impact of insufficient abilities of quality control personnel.

(2) Highlight key quality control and achieve multi-level quality control process. The processing of the simplex zone aims to achieve fidelity and amplitude-preserving, with a focus on strengthening quality control in processing steps such as static correction, denoising, and deconvolution; Multi-work area co-processing aims to achieve the quality of splicing imaging, with a focus on quality control in aspects such as wavelet shaping, residual static correction, and migration velocity pickup.
(3) Strictly monitor the plan and achieve formalized quality control evaluation. The integrated monitoring scheme of point, line, surface, volume, process-interpretation is adopted. Based on enterprise standards and industry guidance, we provide feedback on specific issues in the data to relevant responsible persons in the "Quality Control Analysis and Evaluation Opinion Form", provide rectification direction and opinions, ensuring that there is evidence to support and evidence to investigate.
(4) Use professional software to standardize the data format. We establish a unified data format, and standardize quality control documents, within professional quality control software. Comprehensive quality control content, standardized drawings, one-to-one correspondence between quality control documents and data content, achieving form standardization and ensuring the accuracy and completeness of data.
(5) Strengthen information network cooperation and achieve informationization of data management. Deeply integrating with information, processing process operations, seismic data volumes, and quality control documents are stored together, and a basin level large database is established. Comprehensive information management of seismic data is carried out to provide comprehensive seismic information for exploration and deployment of oilfield companies.

## 4 Example

The post-stack splicing cannot fundamentally solve the problem of inconsistent boundaries, amplitudes, and frequencies in seismic work areas, which affects the improvement of reservoir recognition accuracy. To meet the needs of high-precision structural interpretation and 3D geological modeling at the zone level, high-precision pre-stack merging processing has been carried out in the Gulong area, involving a total of 24 seismic work areas with a full coverage area of 5070 km$^2$ and a data volume of 90TB. It is the largest multi-zone processing task independently undertaken by Daqing Oilfield. The ground elevation of the work area is between 114–182 m. It has a large span of seismic acquisition (1999 to 2022), with significant differences in coverage and energy (Fig. 1); The signal-to-noise ratio of seismic data in some areas (such as Ying 31, Longnan, Ying 88) is low, and there are obvious splicing traces in the post-stack spliced data, which cannot meet the interpretation requirements (Fig. 2).

Using the "Five modernizations of fine quality control" method, the data name and format are unified management (Fig. 3). For SGY format data, SegyInput module is used to unencode and GeoTapeOutput is used to transcribe; The internal format of other software is decoded using the MultiDataInput module, and then transcribed and stored using GeoTapeOutput. The format of the data name is abbreviated as work area + processing unit + year/year + identification + swath. At the same time, the data type marking card is established, which breaks the data format barrier while retaining the complete track header information and processing parameter information.

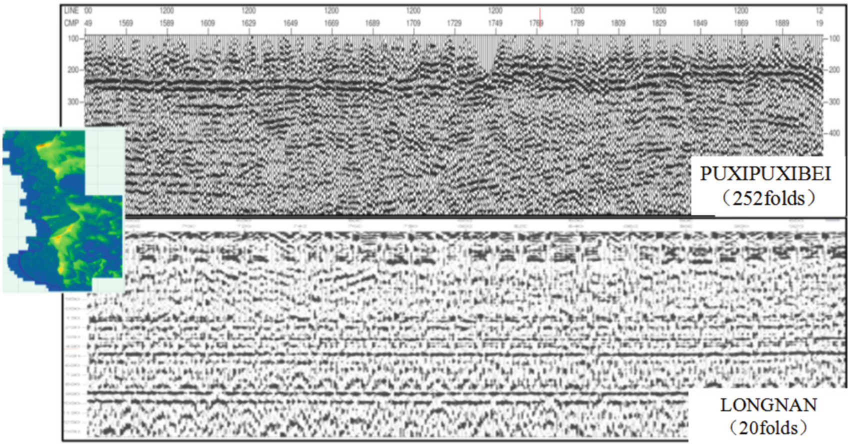


**Fig. 1.** Original stacked profile

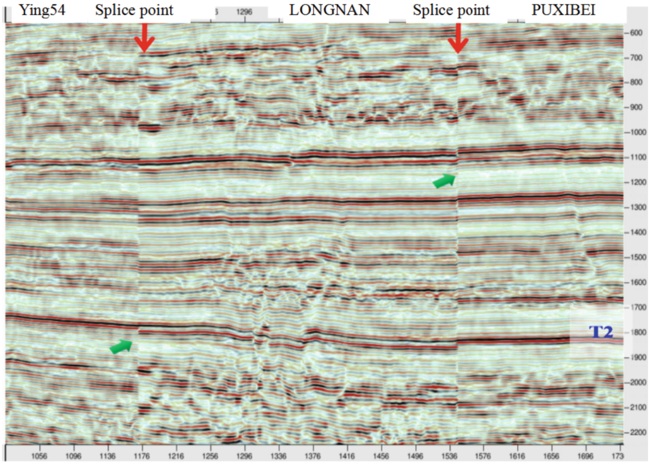
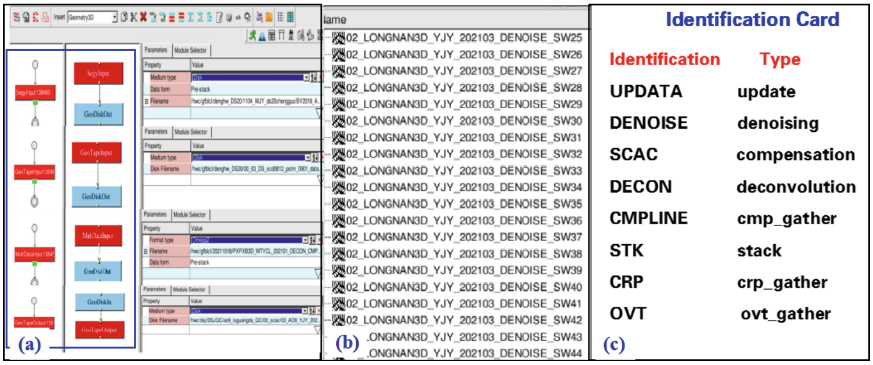


**Fig. 2.** Post-stack splicing profile

After the treatment, the cycle of pre-stack time migration was completed in only 11 months, which needed at least 2–3 year before. The results of multi-zone pre-stack migration eliminate the boundary static correction problem and the energy difference of different periods, and significantly improve the quality of continuous slices compared with post-stack splicing, which lays a data foundation for the subsequent target processing with higher efficiency and higher precision (Fig. 4).



Parameters | Module Selector

| Property | Value |
|---|---|
| Medium type | Disk |
| Data form | Pre-stack |
| Filename | /hwc/gfblcl/denghw_DS20/1104_WJY_ds20chengguo/BY2018_A... |

Parameters | Module Selector

| Property | Value |
|---|---|
| Medium type | Disk |
| Disk Filename | /hwc/gfblcl/denghw_DS20/00_03_DS_scd0812_pstm_0901_data... |

Parameters | Module Selector

| Property | Value |
|---|---|
| Format type | Omega2 |
| Filename | /hwc/gfblcl/20211018/PXPXB3D_WTYCL_202101_DECON_CMP... |
| Data form | Pre-stack |

Parameters | Module Selector

| Property | Value |
|---|---|
| Medium type | Disk |
| Disk Filename | /hwc/dsj/DSJQC/ao9_luguangda_QC/03_scac/03_AO9_YJY_202... |

02_LONGNAN3D_YJY_202103_DENOISE_SW25
02_LONGNAN3D_YJY_202103_DENOISE_SW26
02_LONGNAN3D_YJY_202103_DENOISE_SW27
02_LONGNAN3D_YJY_202103_DENOISE_SW28
02_LONGNAN3D_YJY_202103_DENOISE_SW29
02_LONGNAN3D_YJY_202103_DENOISE_SW30
02_LONGNAN3D_YJY_202103_DENOISE_SW31
02_LONGNAN3D_YJY_202103_DENOISE_SW32
02_LONGNAN3D_YJY_202103_DENOISE_SW33
02_LONGNAN3D_YJY_202103_DENOISE_SW34
02_LONGNAN3D_YJY_202103_DENOISE_SW35
02_LONGNAN3D_YJY_202103_DENOISE_SW36
02_LONGNAN3D_YJY_202103_DENOISE_SW37
02_LONGNAN3D_YJY_202103_DENOISE_SW38
02_LONGNAN3D_YJY_202103_DENOISE_SW39
02_LONGNAN3D_YJY_202103_DENOISE_SW40
02_LONGNAN3D_YJY_202103_DENOISE_SW41
02_LONGNAN3D_YJY_202103_DENOISE_SW42
.ONGNAN3D_YJY_202103_DENOISE_SW43
.ONGNAN3D_YJY_202103_DENOISE_SW44

Identification Card

| Identification | Type |
|---|---|
| UPDATA | update |
| DENOISE | denoising |
| SCAC | compensation |
| DECON | deconvolution |
| CMPLINE | cmp_gather |
| STK | stack |
| CRP | crp_gather |
| OVT | ovt_gather |

**Fig. 3.** (a) Data input template (b) Standardized data naming (c) Identification card

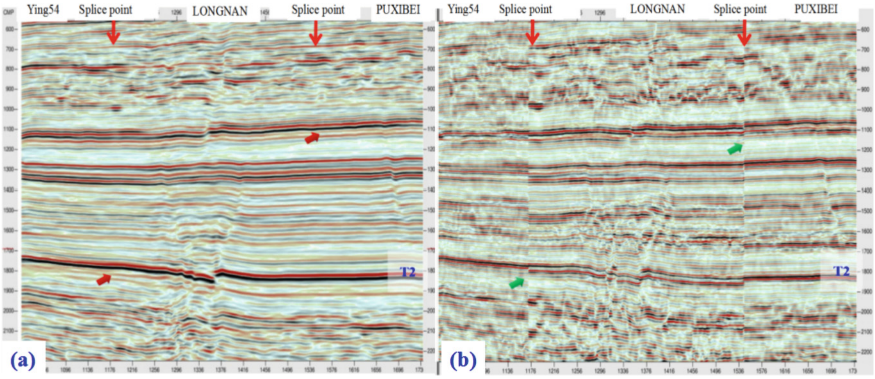


**Fig. 4.** (a) Continuous processing profile of pre-stack migration (b) Poststack splicing profile

## 5 Conclusion

At present, seismic processing has gradually developed from single working area to basin-level multi-zone processing. The “Five modernizations of fine quality control” method has been used to normalize and standardize the management of pre-stack big-data. It meets comprehensive quality control of the basic processing, flow, parameters and effects in multi-zone pre-stack migration processing, by using point, line, surface, volume, and the integrated processing-interpretation means. The comprehensive management of basin-level massive pre-stack data method effectively improves the data quality, provides high-quality intermediate processing results data for basin-level and zonal-level sequential processing, can shorten the processing cycle by 45%–85%, effectively improves the application rate and timeliness rate of seismic results, and realizes the rapid support for exploration and development.

## References

1. Sun, L., Wang, G.: On Daqing Field's revitalization strategy. Daqing Petrol. Geol. Dev. J. **38**(05), 1–7 (2019)
2. Wenyuan, H., Bo, L., Jinyou, Z., et al.: Exploration of geological characteristics and key scientific issues of Gulong shale oil in the Songliao Basin. Earth Sci. J. **48**(01), 49–62 (2013)
3. Wei, X., Bin, L., Xin, L., et al.: Petroleum seismic exploration systems and software development in the era of big data. Sci. Technol. Rev. J. **35**(15), 57–62 (2017)

# Design and Implementation of Cloud-Based Transformation for Traditional Logging Applications

Kun Shao[1,2](✉), Jun Zhou[1,2], Zheng-zhi Zhou[1,2], Xin Chen[1,2], Guo-jun Li[1,2], and Yi-chen Sun[1]

[1] Logging Technology Research Institute of CNPC Logging Co., Ltd., Beijing, China
shaokun.gwdc@cnpc.com.cn

[2] China National Petroleum Corporation Key Laboratory of Well Logging, Xi'an, China

**Abstract.** In the process of digital transformation of logging, how to integrate traditional business applications with digital and intelligent technologies that have been accumulated for many years has become an important topic for research and exploration in the logging industry. The migration of traditional logging applications to the cloud has gradually become an effective means of digital transformation.

China Petroleum Organization has launched the construction of the Exploration and Development Dream Cloud Platform. This article analyzes the characteristics of traditional logging heterogeneous systems and refers to the Dream Cloud technology framework. It classifies and studies the cloud migration modes, proposes three cloud migration modes, namely basic infrastructure migration, thin client migration, and microservice migration, based on the complexity of the architecture of traditional logging applications. It forms a complete set of cloud migration solutions for traditional logging applications.

In response to the characteristics of the logging industry software, breakthroughs have been made in the communication integration of heterogeneous systems, data transparency exchange, and cloud-based data processing, effectively improving the overall service capabilities of specialized logging software.

**Keywords:** Logging application · Microservice · Application Cloudification · Cloud Platform

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20–22 September 2023.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 298–305, 2024.
https://doi.org/10.1007/978-981-97-0272-5_26

# 1 Introduction

With the rapid development and popularization of information technology, new technologies such as cloud computing, big data, and the Internet of Things are widely used in enterprise digital construction and reform. When enterprises practice cloud computing, the entire cloud computing not only plays a huge role in personal data storage, but also in enterprise data sharing and communication. As cloud computing applications deepen, their commercial value is widely recognized, which makes them have an irreplaceable advantage in opening up new application scenarios.

Facing the emerging tide of new information technology, based on the development strategy of the enterprise, China Petroleum Research Institute has explored the technological route of platform development, laying a foundation for the digital transformation and intelligent and stable development of upstream businesses. Currently, it has entered a new stage of shared intelligence [1–5]. System aims to build "Digital China Petroleum Logging," make full use of information and digital technologies, realize new capabilities such as intelligent operations, networked collaboration, and personalized services, and create a new development model driven by users, data, and innovation, to achieve collaborative sharing, continuous innovation, risk control and smart decision-making mechanisms, and constantly improve the labor productivity and asset creation efficiency of all employees. The full cloudification of logging business has become an effective means to achieve this goal.

Logging companies have accumulated a large amount of software resources and logging business asset data in the development and interpretation of logging products in the past 20 years, which are suitable for cloud-based resource integration and data sharing. In the process of building a multidisciplinary and collaborative logging cloud platform and service cloudification, the entire environment needs to be tested synchronously and functionally optimized and perfected to ensure the stable operation of the entire cloud computing environment to support the company's logging interpretation work.

# 2 Design Principles and Architecture

## 2.1 Design Objectives

From the perspective of digital transformation, the cloudification of traditional applications is another form of presentation of existing excellent applications, which enriches the deployment methods and work scenarios of traditional applications. The goal of this design is to transform existing applications through architectural refinement and optimization, refactor redundant modules, and increase rich forms of data exchange to achieve desktop application service, and ultimately achieve application system data interoperability, application interconnection, and result mutual recognition.

Based on the integration of cloud computing related technologies and logging business, a cloud computing and microservice-based application system is built, closely around the development of logging data acquisition, processing, interpretation, evaluation, and other businesses, to create a new logging data processing format with diversified business applications and systemized business processing capabilities. This will achieve centralized resource management, explore new enterprise hybrid cloud construction models to reduce the company's information construction and operation and maintenance costs, enhance core competitiveness, and accelerate digital transformation.

Three major transformations are proposed:

(1) From single well interpretation to multi-well evaluation and reservoir research transformation, to help improve oil and gas evaluation capabilities. Conduct technical research on fine strata comparison, sand body distribution law, main control factors of reservoirs, etc., to achieve accurate calculation of reservoir parameters, evaluation of advantageous reservoirs, recognition of fluid properties, residual oil evaluation, prediction of favorable areas for increasing reserves, etc. Logging can significantly improve the exploration and development guarantee.
(2) From single geological factors to comprehensive geological engineering factor evaluation transformation, to improve reservoir drilling rate and fracturing section accuracy. Conduct technical support for unconventional 3D geological modeling, rapid forward and inverse analysis, geological orientation, horizontal well interpretation and evaluation, production capacity prediction, section clustering and fracturing engineering schemes, exploration deployment, development schemes, etc., significantly improve the lifecycle evaluation capabilities.
(3) From manual and semi-automatic logging processing interpretation to intelligent logging interpretation, to promote the transformation of logging interpretation methods. Conduct research on intelligent logging interpretation technology based on machine learning, achieve intelligent processing, online analysis, intelligent recognition and fast decision-making functions, and realize the intelligent breakthrough of logging interpretation technology and process.

## 2.2 Overall Architecture

Currently, there are many isolated applications in the logging business, and existing systems (products) cannot share data or coordinate business due to existing "barriers". This cannot complete integrated work effectively or respond to user needs efficiently. Therefore, unified standards for data, access, and applications are required. Through software cloudification, application interconnection and integration can be realized to play the overall advantage of software technology. Finally, the cloudized services are integrated into the logging cloud computing application system, serving the logging big data platform and related ecological applications. The system is built on one set of basic infrastructure platform + one set of support platform + N oil and gas business scenarios. The microservice architecture is adopted to continuously provide capabilities and have high scalability by reasonably splitting and serving the functions. The architecture of the entire cloud application system is as follows (see Fig. 1):

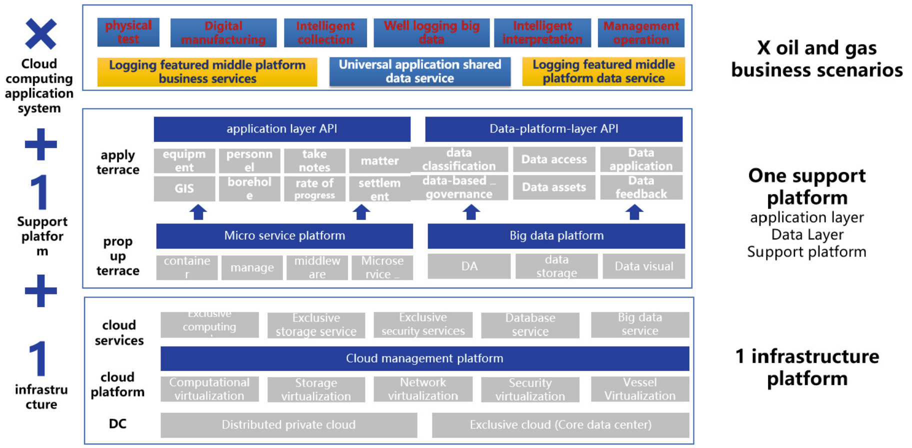


**Fig. 1.** 1 set of infrastructure platform + 1 set of support platform + N oil and gas business scenarios as the basis for construction. The microservices architecture is used to enable the system to provide continuous capability with a high degree of scalability by splitting and servicing functions in a logical manner.

**Data Architecture.** The data architecture for logging business and interpretation and evaluation mainly includes logging business systems and data middleware. The most important business data is stored in various business systems and synchronized to the data middleware. The operations center obtains data from the data middleware for processing, and forms data storage into the analysis layer of the data middleware, which integrates the portals and permissions of various business systems.

**Technology Architecture.** Based on the logging cloud platform and relying on the data middleware, the system adopts a distributed computing technology and microservice architecture, following the overall technical plan of "one platform, one system, multiple scenarios, and micro-applications". The system is developed using SpringBoot, Mybatis, Bwuap_Security, and Freemarker, and data interaction and sharing are achieved through the data middleware.

## 3 Key Technology Implementation

### 3.1 Logging Software Cloudification Strategy

Most existing logging acquisition, processing, and interpretation software is mainly based on monolithic applications and is heavily coupled internally, requiring them to be split and decoupled for cloudification. Relevant solutions should be designed from two dimensions: one is the complexity of monolithic applications, and the other is the technical approach to cloudification. The complexity of monolithic applications can be divided into four categories based on the modules of traditional business application structure, which are:

- Data Access

- Data Access + Data Processing
- Data Access + Data Processing + Data Presentation
- Data Access + Data Processing + Data Presentation + Interactive Operations

Based on this, relevant applications can be classified by complexity:

- Complex System (Data Access + Data Processing + Data Presentation + Complex Interactive Operations)
- Moderately Complex Application (Data Access + Data Processing + Data Presentation + Simple Interactive Operations)
- Lightly Coupled Simple Application (Data Access + Data Processing)
- Uncoupled Application or Method (Data Access + Data Processing)

**Table 1.** Different applications correspond to the cloud solution

| | Could Desktop | Micro services | Application Services |
|---|---|---|---|
| Complex System | √ | | |
| Moderately Complex Application | | | √ |
| Lightly Coupled Simple Application | | √ | |
| Uncoupled Application or Method | | √ | |

Implementation of lightly coupled simple applications (data access + data processing) and decoupled applications or methods (data access + data processing):

- Data access module + data processing module encapsulation = 1 micro service
- Data access micro service + data processing micro service = 2 micro services
- Data access micro service + algorithm micro service + data processing logic encapsulation micro service for module A = 3 micro services

For moderately complex applications (data access + data processing + data presentation + simple interactive operations):

- Not split into micro service form, published as a web service or Windows Server
- Module data access + online independent application management + offline independent application management + local independent data application = cloud publishing mode

For complex systems (data access + data processing + data presentation + complex interactive operations):

Data access module + complex system = cloud desktop.

### 3.2 Logging Plot Cloudification Technology

The logging plot module is a highly complex and widely used function. Both logging collection and logging interpretation have accumulated a large number of application templates to meet the needs of oilfield applications. Cloudifying the logging plot desktop application makes it possible to fully present it on the web and has significant implications for logging cloud applications. In the practical process, the system combines the original logging plot technology with Google map tiling technology to create an HTML5-based online logging curve drawing technology.

Due to the high complexity of the plotting module, according to the cloudification strategy, the plotting module of the existing desktop application needs to be decoupled to form an independent application, and then secondary encapsulation is achieved through self-developed RPC transformation technology to realize server-side drawing services. The logging curve graph is drawn on the server side, and for some larger imaging logging graphs, they are cut into pieces to achieve a smooth visual effect. Data files are cut into fixed size blocks, which can be set before transmission according to the network conditions. Data cutting divides the complete graph into many slices, which are transmitted to the client and then spliced into a complete image on the client side. There are two ways to splice data: real-time acquisition and interpretation browsing. In real-time acquisition mode, the images are spliced by appending them according to the transmission order of the images. In the interpretation browsing mode, data splicing is done in a filling manner.

In order to further save bandwidth, the curve image data obtained from the server side can be buffered to the disk of the client machine. When the curve display enters the interpretation browsing mode and the user drags the scroll bar to a specific location, data loading has just begun. The browser side sends a request for the specified depth curve to the server to obtain the curve data, which will be cached. If the user drags the scroll bar to that position again next time, the curve information will be directly read from the client instead of requesting corresponding data from the server again, to reduce the unnecessary bandwidth consumption during transmission.

Currently supported logging data are divided into depth-indexed data and time-indexed data. Among them, depth-indexed data is divided into uphole logging and downhole logging, which are only distinguished in real-time display. During the logging process, when logging starts, it automatically switches to depth mode or time mode according to the logging method. During interpretation browsing, when a logging data file is opened, it automatically switches to depth mode or time mode according to the index method of the file.

### 3.3 Cloudification Technology for Logging Data Processing Modules

The operation of logging data processing modules requires the support of underlying resources such as CPU, disk, and network. Containers can easily and quickly package and run modules on physical or virtual infrastructure, and are relatively lightweight compared to virtual machines, while also improving resource utilization of the underlying infrastructure. With the fluctuating demand, containers can dynamically start or stop application instances and easily migrate them to different environments.

At the same time, the operation of logging data processing modules is real-time, and the system cannot predict which logging data processing modules need to run at the same time, so container cluster management tools are very suitable for dynamic runtime management of logging data processing modules.

Container cluster management tools can manage multiple containers combined into an application on a group of servers. Each application cluster is seen as a deployment or management entity in the container orchestration tool. Container cluster management tools provide automation for application cluster management, including application instance deployment, application updates, health checks, elastic scaling, automatic fault tolerance, etc.

In practice, logging data processing modules are encapsulated into dynamic link libraries, and server applications are formed through self-developed RPC to place these application services in containers for dynamic management, forming a set of container-based logging data processing module invocation technology.

## 4 Application Effect

The logging cloudification transformation technology has been fully integrated into the Logging Cloud Computing Application System, and related modules have also supported the function implementation of multiple web application systems in the logging big data platform ecology.

### 4.1 Application of Logging Cloud Visualization Components

Currently, the logging cloud visualization components have been applied to integrated application systems of oilfields in Xinjiang, Dagang, Qinghai, Jilin, Southwest, etc., providing online display function of logging data for Logging's regional branches, ensuring that the display content complies with relevant technical specifications and meets on-site application requirements. Compared to traditional client-side logging curves, the HTML5-based online logging curves plotting method is flexible and targeted, allowing the logging curves to be plotted in the browser without any dependencies and fully adapted for mobile display, while significantly reducing performance requirements on the browser side due to the back-end plotting method and improving the web display of high-end imaging data from logging.

### 4.2 Application of Logging Data Processing Technology

The cloudification technology for data processing has significant effects on wellbore quality and cementing quality cloud processing. In order to cooperate with the group company's three-year centralized rectification plan for oil, gas and water wells, according to the group company's "Red Line for Judging Unqualified Wellbore Quality and Cementing Quality", the wellbore and cementing quality analysis data processing algorithms were cloudified to form a logging wellbore quality analysis system. This system greatly improves the efficiency and quality of the centralized rectification plan.

## 5 Conclusion

Based on the demand for traditional logging applications to be implemented in the cloud, this paper classified traditional applications and proposed three cloud migration solutions. The key technologies involved in the solutions were implemented, and logging business software was combined with excellent internet technologies in key areas such as logging graph cloudification and processing module cloudification. Targeted technical solutions were proposed to reduce the difficulty of development, testing, and implementation of traditional applications in the cloud. In the digital construction process, these relevant technologies have been iteratively applied, and it is believed that the collision of such technologies will generate more excellent technologies in the logging field, thereby improving the overall application capability of logging business software.

## References

1. Jia, A., Guo, J.: Key technologies and understandings on the construction of smart fields. Petrol. Explor. Dev. **39**(1), 118–122 (2012)
2. Gao, C., He, J., Huang, Z., Liu, G., Fang, C., Pan, W.: Digital basin: a new stage for studying the Chinese petroliferous basins. Petrol. Explor. Dev. **31**(5), 433–139 (2009)
3. Shi, Y.: Analysis of the research status of intelligent oilfield in China. Technol. Ind. Across Straits (12), 81–83 (2016)
4. Peng, C.: Research and thinking on top-level design of enterprise informatization China management informationization **17**(10), 40–41 (2014)
5. Jianfeng, L.: Development and problems of digital oil field in China. Digit. Chem. Eng. **9**, 4–7 (2004)
6. Li, C.: Event driven architecture and application [R/OL]. (2009-03-16) [2020-07-20]. http://www.uml.org.cn/soa/201001295.asp
7. Ma, T., Xu, Z., Wang, T., et al.: Research on software architecture of digital oilfield. Inf. Technol. Inf. (6), 41–45 (2010)
8. Xu, Z., Ma, T., Wang, T., et al.: Exploration and discussion on the development of digital oilfield technology. Formation Chma (9), 28–32 (2012)

# Research on Fault Diagnosis System of Key Drilling Equipment Based on Internet of Things

Xue-li Luo[1,2], Yi Jin[1], Xiao-guang Yang[1], Deng Jia[2](✉), Yong Su[1], Yi Zhang[1], Yong-chao Wang[1], Bing-deng Chen[1], and Han-qin Bai[1]

[1] CNPC Engineering Technology R&D Company Limited, Beijing, China
[2] Beijing Kembl Petroleum Technology Co. Ltd., Beijing, China
jiadeng123@qq.com

**Abstract.** As the important equipment in the oil drilling rig system, the operation status of drilling pump and winch directly affects the safety and efficiency of oilfield drilling production. The existing inspection and maintenance of drilling equipment mainly rely on manual patrol inspection and post-maintenance. The intelligent monitoring and health diagnosis technology of drilling key equipment can help to realize the life cycle management of drilling key equipment and significantly improve the level of drilling equipment evaluation business. For the health monitoring and fault diagnosis analysis of drilling pump and winch, this paper focuses on the research and design of the online monitoring and fault diagnosis system of drilling pump and winch using the Internet of Things technology. This system can realize remote real-time monitoring and fault diagnosis of the drilling process, reduce the workload of on-site personnel, improve the management efficiency of equipment, and more safely ensure the exploration and development of oil and gas resources.

**Keywords:** Drilling rig · Condition monitoring · Data acquisition · Fault diagnosis

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 306–318, 2024.
https://doi.org/10.1007/978-981-97-0272-5_27

## 1 Introduction

There are many inducements of oil drilling rig site environment, complex operation flow, safety risks and equipment failure, which affect the safe operation of drilling rig and the normal work of on-site testing instruments. Therefore, all-weather condition monitoring and fault early warning of drilling rig key equipment is a necessary measure to ensure its safe and stable operation [1, 2]· In order to accurately monitor and diagnose the running condition of the key equipment of the drilling rig, through the comparison and investigation of the technologies commonly used in rotary motion and reciprocating equipment monitoring at home and abroad, different sensors are used to monitor the vibration signal, temperature, pressure, voltage, current, frequency and other signals of the equipment in real time, and the vibration sensor is used to monitor and analyze the vibration and temperature at the bearing of drilling pump and winch equipment. The equipment failure can be predicted in advance to ensure the normal operation of the equipment [3, 4]. This paper first analyzes the monitoring mode of drilling pump and winch, and then analyzes the common faults of drilling pump and winch and the layout of monitoring points, as well as the state parameters that need to be monitored. Finally, according to the characteristics of parameters, equipment structure and field environment, the software and hardware system for on-line monitoring and fault diagnosis of drilling pump and winch is designed, which can effectively monitor the operation status of key equipment in in-service and remote drilling rigs and predict the occurrence of equipment faults. Prevent major accidents in the key equipment of the drilling rig from affecting the normal drilling operation.

## 2 The Choice of Monitoring Mode

Monitoring technology can be divided into two categories, off-line monitoring and on-line monitoring. Offline monitoring, also known as regular monitoring. The on-the-spot data are collected regularly or irregularly, analyzed and diagnosed on the spot, or put back to the computer for monitoring and diagnosis analysis by computer software. The advantage is that it can be analyzed carefully, and the disadvantage is that it is powerless to deal with sudden failures. Online monitoring, also known as online diagnosis. The mechanical equipment fault diagnosis system is connected with the tested equipment, which can monitor the current state of the equipment in real time, catch sudden faults and analyze them in time, which is also the advantage of on-line diagnosis [5].

This paper adopts an on-line monitoring system, which fully considers the working characteristics of drilling pump and winch. By monitoring the vibration signal, temperature and pressure of drilling pump and winch, and storing the monitoring data to the server in real time through the network, any customer with authority can view the data through Internet, which can easily understand the running status of the equipment, carry out fault analysis, and provide a basis for operation or maintenance.

# 3 Common Fault Analysis and Measuring Point Layout of Drilling Pump and Winch

## 3.1 Fault Analysis of Drilling Pump and Winch

The main function of the drilling pump is to cool the bit temperature by injecting some flushing media from the ground surface, including flushed mud, water and flushing fluid, through faucets and pressure pipes to the central hole of the drill string by giving appropriate external pressure. Achieve bit temperature cooling and pump all kinds of debris produced during drilling to the surface [6–8]. Taking the QDP- 2200 five-cylinder pump as an example, the main components of the drilling pump include air bag, pump head, piston rod, suction pipe, connecting rod, crankshaft, crosshead and discharge pipe. The main vulnerable parts of drilling pump are piston, cylinder liner, plunger, seal and pump valve. The overall structure of the drilling pump is shown in Fig. 1.

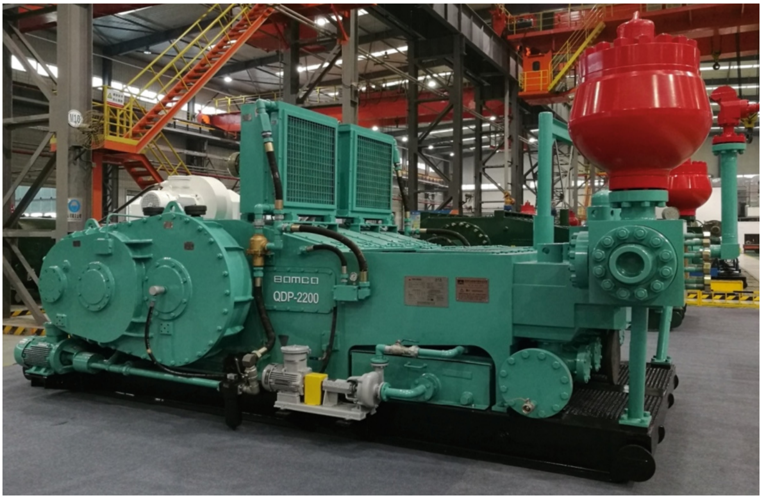


**Fig. 1.** Overall structure diagram of QDP- 2200 five-cylinder pump

The drilling winch is not only the hoisting system, but also the core component of the whole drilling rig. Taking the JC70DB drilling direct drive winch as an example, the winch consists of AC variable frequency motor, shaft coupling and drum. The drum is mainly composed of cylinder shell, blessing plate, wheel Yi, bearing housing, drum shaft, etc. [9, 10]. The drum is connected with the driving device, the driving device is outside the drum, and the drum output shaft is connected with the coupling, which belongs to the externally driven drum. As an important transmission part of the winch, not only the structure is complex, but also the load is more complex, it is the most common damaged part of the winch system. The overall structure of the winch is shown in Fig. 2.

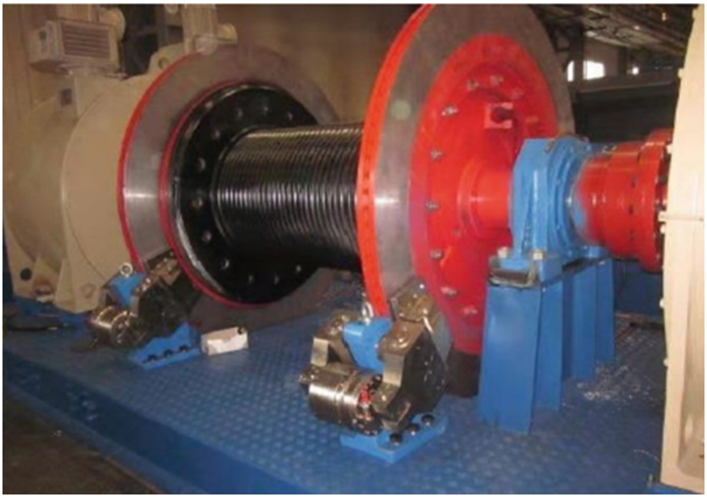


**Fig. 2.** Overall structure diagram of JC70DB winch

By analyzing and comparing the structure composition, working principle and operating environment of drilling pump and winch equipment, the faults of drilling pump and winch are summarized as shown in Table 1 and Table 2.

**Table 1.** Common faults of drilling pump

| Basic structure | | Main components | Fault classification |
|---|---|---|---|
| Hydraulic end | Piston assembly | Piston ring | Wear, fracture, leakage |
| | | Support ring | Wear, fracture |
| | | Bolt | Loose |
| | | piston rod | Wear, fracture |
| | Cylinder block | Liquid cylinder | Pull cylinder, Bump cylinder |
| | | Stuffing box | Packing wear |
| | | Suction valve | Leakage, valve plate fracture |
| | | Exhaust valve | Leakage, valve plate fracture |
| Power end | Motor | Rolling bearing | Fatigue shedding, wear, deformation, corrosion, fracture |
| | | Rotor shafting | Imbalance, misalignment |
| | Shaft coupling | | Misalignment |

(*continued*)

**Table 1.** (*continued*)

| Basic structure | | Main components | Fault classification |
|---|---|---|---|
| | Crankshaft | Crankshaft body | Imbalance, misalignment |
| | | Rolling bearing | Fatigue shedding, wear, deformation, corrosion, fracture |
| | Medium rod | | Wear and tear |
| | Connecting rod structure | Big head tile | Wear and tear |
| | | Connecting rod bolt | Loose, fracture |
| | | Small head tile | Wear and tear |
| | Crosshead | Crosshead pin | Wear, fracture, loose |
| | | Up and down slide | |
| | Gear reduction box | Rolling bearing | Fatigue shedding, wear, deformation, corrosion, fracture |
| | | Gear | Wear, broken teeth |

**Table 2.** Common faults of winch

| Basic structure | Main components | Fault classification |
|---|---|---|
| AC variable frequency motor | Rolling bearing | Fatigue shedding, wear, deformation, corrosion, fracture |
| | Rotor shafting | Imbalance, misalignment |
| Shaft coupling | | Misalignment |
| Roller shaft assembly | Roller body | Imbalance, misalignment |
| | Brake disc | Wobble, wear, abnormal noise |
| | Rolling bearing | Fatigue shedding, wear, deformation, corrosion, fracture |

### 3.2 Layout of Vibration Monitoring Points for Drilling Pump and Winch

According to the structural characteristics of drilling pump and the installation standard of vibration sensor, 10 vibration acceleration sensors are mainly arranged at the hydraulic end (suction valve and discharge valve). The connecting rod crosshead is mainly equipped with 5 vibration acceleration sensors, 5 displacement vibration sensors, 1 key phase sensor, 2 vibration acceleration sensors in the crankcase and 4 vibration acceleration sensors in the gear reducer. The motor is equipped with four vibration accelerometers. The specific layout of the measuring points is shown in Fig. 3.

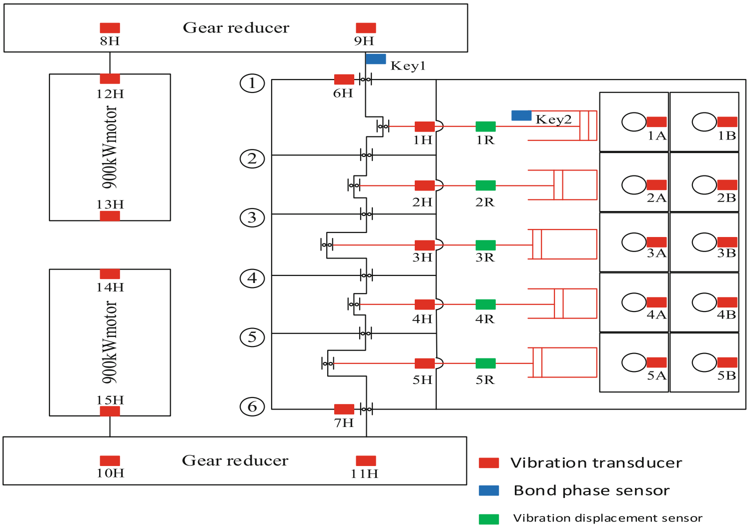


**Fig. 3.** Layout of monitoring points for drilling pump

According to the structural characteristics of the winch and the installation standard of the vibration sensor, four vibration acceleration sensors are arranged in the No. 1 motor, four vibration acceleration sensors are arranged in the No. 2 motor, and four vibration acceleration sensors are arranged in the winch bearing. The specific layout of the measuring points is shown in Fig. 4.

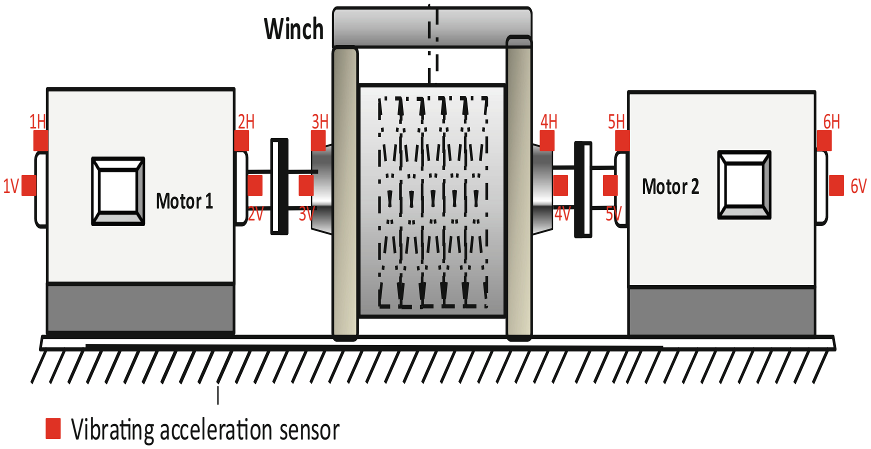


**Fig. 4.** Layout of monitoring points for winch

# 4 Design of On-Line Monitoring and Diagnosis System for Key Equipment of Drilling Rig

The network architecture of online monitoring and diagnosis system for key equipment of drilling rig mainly includes two parts: hardware and software. The hardware mainly includes a vibration temperature sensor, a gateway signal collector and a server, which are used to obtain real-time vibration and temperature signals of drilling pumps and winches, and transmit the signals from the scene to the workstation in the central control room [11–14]. The software includes data acquisition software and fault diagnosis software. The software module mainly includes six modules: on-line monitoring, intelligent diagnosis, prediction and maintenance, historical trend, equipment information and parameter setting. It is used for real-time condition monitoring, typical fault diagnosis and vibration trend analysis of drilling pump and winch. The whole system consists of vibration temperature sensor, data acquisition and processing unit, gateway signal collector, Ethernet network, server, database software and analysis software, and the data can be shared remotely through the Internet. The system composition block diagram is shown in Fig. 5.

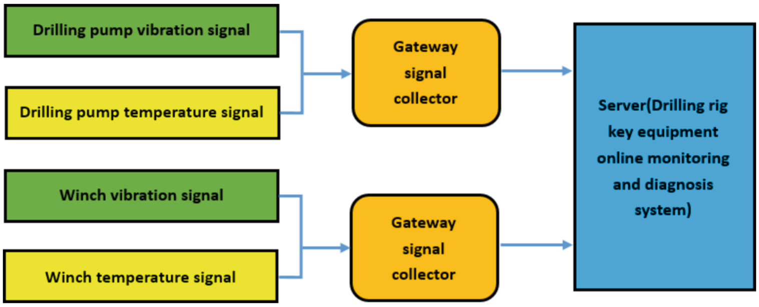

Drilling pump vibration signal
Drilling pump temperature signal
Winch vibration signal
Winch temperature signal
Gateway signal collector
Gateway signal collector
Server(Drilling rig key equipment online monitoring and diagnosis system)

**Fig. 5.** Block diagram of on-line monitoring and diagnosis system for key equipment of drilling rig

## 4.1 Hardware Design and Implementation

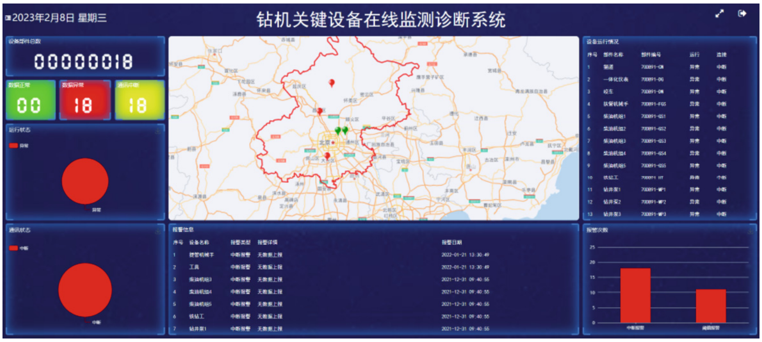
The hardware part is mainly composed of vibration and temperature sensor, gateway signal collector and server industrial computer. It mainly realizes the functions of signal acquisition, transmission and storage.

(1) In this system, ZD-530 acceleration sensor is developed to collect vibration and temperature signals. ZD-530 acceleration sensor is a composite sensor for vibration and temperature parameter output. It adopts stainless steel shell, laser welding seal, and double-layer screen has strong anti-interference ability, which can adapt to long-term and reliable operation in harsh environment. It is convenient for monitoring the vibration and temperature parameters of gears, bearings and transmission shafts on rotating machinery.

(2) The system is equipped with a gateway signal collector for data transmission and communication through GPRS/CDMA, 4G mobile signal or LAN interface protocol. The gateway collector is arranged outdoors and connected with all sensors through cables. The enclosure is explosion-proof and can adapt to various harsh environments. A total of 5 gateway signal collectors are installed on the site, and each gateway signal collector can connect 8 sensors.
(3) The monitoring server is an industrial computer with stable performance. The configuration parameters of the server are i5-8265 CPU, 16G memory, 2 pieces of 512G SSD solid-state hard disk, with RS-232 standard communication interface, 4G network card, Ethernet card and WIFI interface. The monitoring host and the gateway signal collector are equipped with a fixed IP to facilitate data transmission with the vibration and temperature sensor acquisition module.

## 4.2 Software Design and Implementation

The on-line monitoring and diagnosis system of drilling rig key equipment carries out corresponding maintenance and maintenance in advance based on the diagnosis results, which can effectively avoid and prevent the failure of drilling rig key equipment and improve the efficiency and integrity rate of drilling rig key equipment [2]. The software is written by using Visual Studio software platform. The program has the functions of data acquisition, on-line monitoring, data analysis, fault diagnosis, historical trend, parameter setting, equipment information query, etc., and realizes the functions of data acquisition, on-line monitoring and fault identification of key equipment of drilling rig. The software interface is shown in Fig. 6. The software system framework is shown in Fig. 7. The functions of each module are shown in Table 3.



设备运行情况

| 序号 | 部件名称 | 部件编号 | 运行 | 连接 |
|---|---|---|---|---|
| 1 | 猫道 | 700891-CW | 异常 | 中断 |
| 2 | 一体化仪表 | 700891-DG | 异常 | 中断 |
| 3 | 绞车 | 700891-DW | 异常 | 中断 |
| 4 | 扶管机械手 | 700891-FGS | 异常 | 中断 |
| 5 | 柴油机组1 | 700891-GS1 | 异常 | 中断 |
| 6 | 柴油机组2 | 700891-GS2 | 异常 | 中断 |
| 7 | 柴油机组3 | 700891-GS3 | 异常 | 中断 |
| 8 | 柴油机组4 | 700891-GS4 | 异常 | 中断 |
| 9 | 柴油机组5 | 700891-GS5 | 异常 | 中断 |
| 10 | 铁钻工 | 700891-IIT | 异常 | 中断 |
| 11 | 钻井泵1 | 700891-MP1 | 异常 | 中断 |
| 12 | 钻井泵2 | 700891-MP2 | 异常 | 中断 |
| 13 | 钻井泵3 | 700891-MP3 | 异常 | 中断 |

报警信息

| 序号 | 设备名称 | 报警类型 | 报警详情 | 报警日期 |
|---|---|---|---|---|
| 1 | 扶管机械手 | 中断报警 | 无数据上报 | 2022-01-21 13:30:49 |
| 2 | 工具 | 中断报警 | 无数据上报 | 2022-01-21 13:30:49 |
| 3 | 柴油机组3 | 中断报警 | 无数据上报 | 2021-12-31 09:40:55 |
| 4 | 柴油机组4 | 中断报警 | 无数据上报 | 2021-12-31 09:40:55 |
| 5 | 柴油机组5 | 中断报警 | 无数据上报 | 2021-12-31 09:40:55 |
| 6 | 铁钻工 | 中断报警 | 无数据上报 | 2021-12-31 09:40:55 |
| 7 | 钻井泵1 | 中断报警 | 无数据上报 | 2021-12-31 09:40:55 |

**Fig. 6.** The software interface

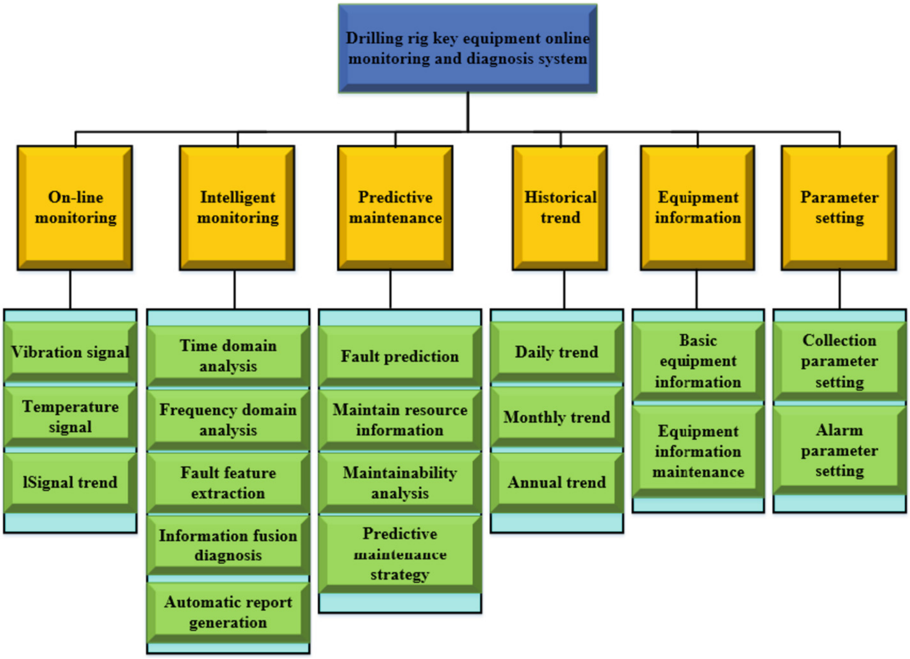


**Fig. 7.** Software system framework diagram

**Table 3.** Function of each module

| Module name | Main function |
|---|---|
| Data acquisition | 1. Read data from the gateway through modbus and tcp interface protocols<br>2. Save the read data in local and cloud databases and display the data<br>3. Calculate the characteristic value, daily average value and monthly average value of the waveform data of each measuring point, and save them in the database |
| On-line monitoring | 1. Display the vibration temperature data collected by the sensor at each measuring point<br>2. When the data of a certain point or some points is greater than the standard value, the corresponding test points will show yellow (high report) or red (high report) to alarm |
| Data analysis | 1. The vibration data of different measuring points at different times can be analyzed in time domain and frequency domain, and the features can be extracted<br>2. According to the selected start and end time to view the time-frequency domain diagram of each measuring point, each time-frequency domain picture can be saved locally for easy viewing |

(*continued*)

**Table 3.** (*continued*)

| Module name | Main function |
|---|---|
| Fault diagnosis | 1. According to the collected data and standard parameters, the fault diagnosis of each measuring point component can be carried out<br>2. The monitoring and diagnosis equipment can be evaluated as a whole<br>3. Generate equipment fault diagnosis report, display the diagnosis results of each measuring point and the overall status of the equipment<br>4. The diagnostic report can be downloaded to the local for viewing as required |
| Historical trend | 1. Historical trend chart and historical data of each measuring point can be viewed |
| Equipment information | 1. Query the basic information about the current equipment, such as model, power rating, etc<br>2. Modify the information content in the corresponding position of the equipment related information, and click the Save button to save the basic information of the equipment |
| Parameter setting | 1. View the current standard value of each measurement point<br>2. View the average value of each measuring point in the last day and the average value of the last month<br>3. Enter the value in the text box of the corresponding measuring point and click the change button to change the current standard value |

## 5 Field Test

In order to verify the accuracy and reliability of the online monitoring and diagnosis system for the key equipment of the drilling rig, the system is deployed on the automatic drilling rig of a company for condition monitoring and fault diagnosis. The QDP-2200 drilling pump is equipped with 31 measuring points to install integrated vibration and temperature sensors, and the JC70DB winch is equipped with 12 measuring points to install sensors. The measuring point diagram of drilling pump 1# high-pressure gun is shown in Fig. 8, and the measuring point diagram of the 6# drainage valve of the drilling pump is shown in Fig. 9.

During the test period, a total of 108 days of monitoring data were collected, and the software system generated 108 drilling pump monitoring reports and 108 winch monitoring reports according to the collected vibration waveforms, including 1024 diagnosis results of drilling pump testing points. After verification, 914 winches were correctly diagnosed, and the accuracy of drilling pump fault identification was about 89.3%. There were 494 winch test points, of which 429 were correctly diagnosed by verification. The accuracy of winch fault identification was about 86.8%. The experimental results show that the system runs smoothly, the data transmission is normal, and the software and hardware design meets the requirements of on-site acquisition. The original record report of the drilling pump is shown in Fig. 10.

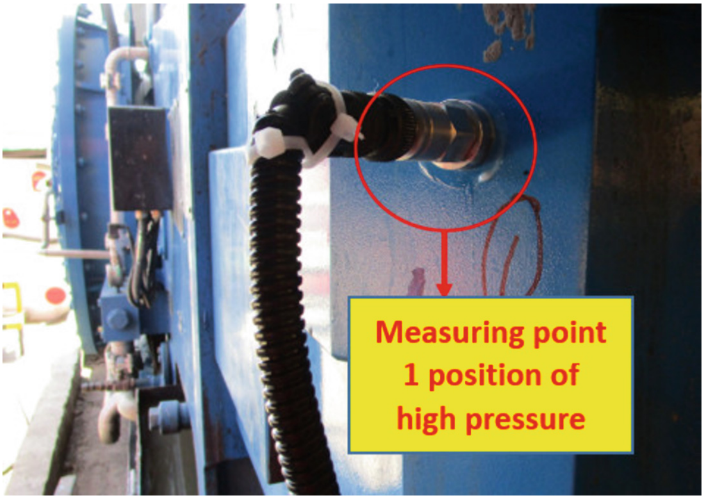


**Fig. 8.** Measuring point 1 position of high pressure

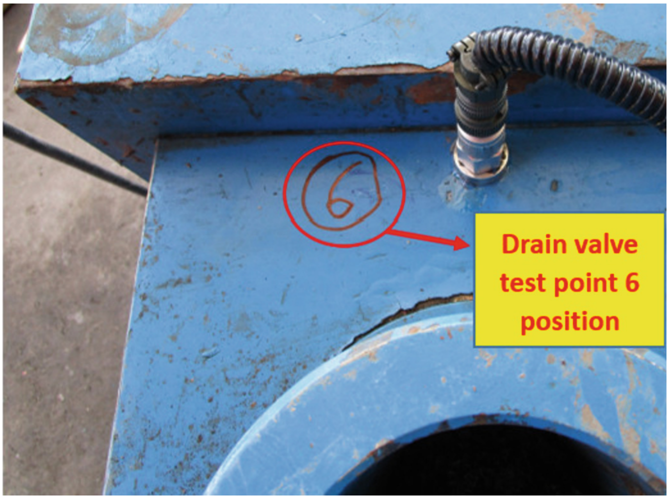
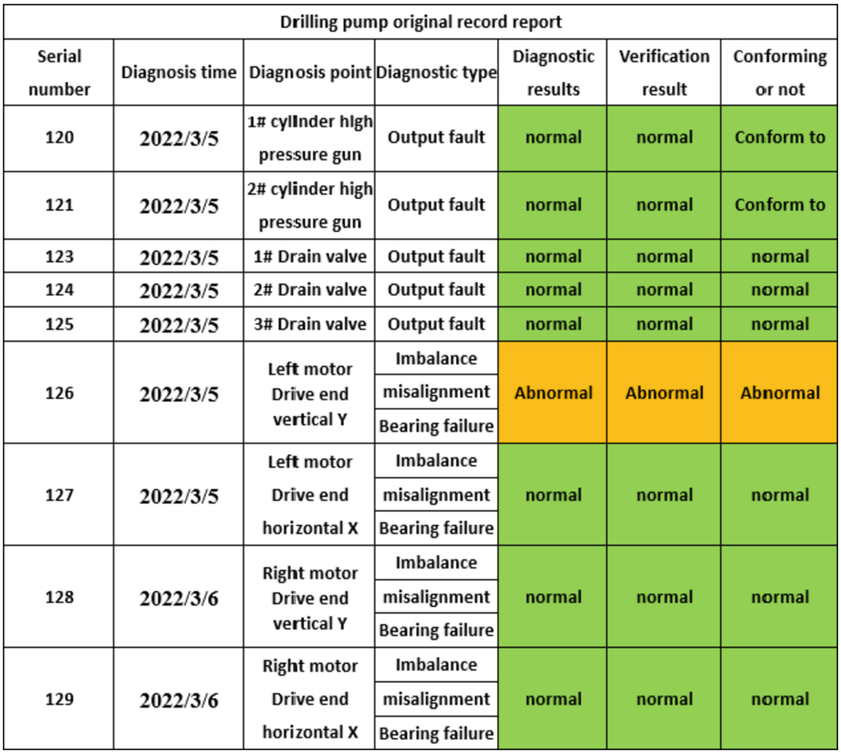


**Fig. 9.** Drain valve test point 6 position

Fault diagnosis system judgment error and unidentified situation further analysis and research. Drilling pump and winch have complex structure, and there are many excitation sources inside. Due to the complex structure, many vulnerable parts, shafting vibration and connecting rod vibration are often caused by the installation deviation of structure and sensor and complex operation conditions. The signal that produces vibration is complex and difficult to deal with, so comprehensive factors should be considered in analysis. In order to further improve the accuracy and reliability of fault diagnosis testing, the following four ways can be used for optimization and improvement:



| Drilling pump original record report | | | | | | |
|---|---|---|---|---|---|---|
| Serial number | Diagnosis time | Diagnosis point | Diagnostic type | Diagnostic results | Verification result | Conforming or not |
| 120 | 2022/3/5 | 1# cylinder high pressure gun | Output fault | normal | normal | Conform to |
| 121 | 2022/3/5 | 2# cylinder high pressure gun | Output fault | normal | normal | Conform to |
| 123 | 2022/3/5 | 1# Drain valve | Output fault | normal | normal | normal |
| 124 | 2022/3/5 | 2# Drain valve | Output fault | normal | normal | normal |
| 125 | 2022/3/5 | 3# Drain valve | Output fault | normal | normal | normal |
| 126 | 2022/3/5 | Left motor Drive end vertical Y | Imbalance<br>misalignment<br>Bearing failure | Abnormal | Abnormal | Abnormal |
| 127 | 2022/3/5 | Left motor Drive end horizontal X | Imbalance<br>misalignment<br>Bearing failure | normal | normal | normal |
| 128 | 2022/3/6 | Right motor Drive end vertical Y | Imbalance<br>misalignment<br>Bearing failure | normal | normal | normal |
| 129 | 2022/3/6 | Right motor Drive end horizontal X | Imbalance<br>misalignment<br>Bearing failure | normal | normal | normal |

**Fig. 10.** Drilling pump original record report

(1) regular inspection of the testing sensor to determine whether the function is normal and whether the layout is loose; (2) to add multiple sensor measurements; (3) to use reliable transmission cables to enhance anti-interference. (4) optimize the deep learning model to improve the ability of fault feature recognition and information fusion.

## 6 Conclusion

Through the analysis and study of the common faults of drilling pumps and winches used in the oil field, in order to obtain the operating parameters of drilling pumps and winches accurately, quickly and reliably, and to improve the integrity of automatic drilling rigs, in this paper, the on-line monitoring and fault diagnosis system of drilling pump and winch is developed based on the Internet of things technology. The system has the functions of data acquisition, recording, cloud upload, online real-time diagnosis, early warning and so on. The functions of data acquisition, on-line monitoring and fault identification of key equipment of drilling rig are realized.

In addition, the condition monitoring test of the drilling pump and winch of the automatic drilling rig is carried out, combined with the time domain map of the horizontal vibration of the measuring point of the vibration signal, the effective value of temperature and vibration intensity of the measuring point are obtained, and the location of the fault is speculated. By comparing with the results of on-site investigation of drilling pump and winch, it is proved that the system is effective for on-line monitoring and fault diagnosis of key equipment of drilling rig, and the accuracy of fault identification of drilling pump is about 89.3%. The accuracy of winch fault identification is about 86.8%. In order to improve the accuracy and reliability of the fault diagnosis system, optimization and improvement analysis and research are carried out. The research results realize the fault early warning and post-event support of drilling equipment, provide strong support for the development of drilling equipment fault diagnosis and monitoring and evaluation business and ensure the safety of oil and gas production.

In the next step of research, the monitoring image video can be added to the fault signal analysis to improve the fault identification rate, and the development of wireless vibration data acquisition module can also be considered to solve the difficulty of arranging cables on the spot.

**Acknowledgments.** The project is supported by Scientific Research and Technological Development Project of CNPC (Number 2021DJ4304).

## References

1. Geng, F.: Design of draw works fault monitoring and remote diagnosis system based on cloud computing, pp. 7–9. China University of Petroleum, Beijing (2021)
2. Hong, X., Duan, L., Yang, X., et al.: Review on the application of intelligent optimization algorithms in mechanical fault diagnosis. Measur. Control Technol. **40**(7), 1–8 (2021)
3. He, Q., Ma, G., Zhang, H., et al.: Remote monitoring and diagnosis system for drilling rig based on cloud platform. China Pet. Mach. **47**(8), 47–54 (2019)
4. Cong, W., Zhang, P., Lin, Z., et al.: Research on the system of drilling rig remote online monitoring and fault diagnosis. China Pet. Mach. **38**(6), 26–30 (2012)
5. Li, T.: The application research of the distributed network monitoring in drilling rig, pp. 36–41. Xi'an Shiyou University, Xian (2015)
6. Zhang, M., Heng, X., Wang, X., et al.: Design of downhole integrated monitoring system for oil-gas wells. China Pet. Mach. **46**(10), 92–95 (2018)
7. Chuan, H., Jing, Y., Xiao, F.: Design of safety early warning system based on multi-sensor data fusion. Appl. Mech. Mater. **3365**(602) (2014)
8. Zhang, B., Gao, X., Li, X.: Complete simulation and fault diagnosis of sucker-rod pumping. SPE Prod. Oper. **36**(02), 277–290 (2021)
9. Tanabe, S., Prasertsintu, T.: Integration of wireless and remote monitoring in condition monitoring systems for offshore application. In: Offshore Technology Conference, 17–19 August, Kuala Lumpur, Malaysia (2020)
10. Wang, H.: Fault prediction and health management of large compressor unit. Chem. Eng. Des. Commun. **43**(4), 103 (2017)
11. Jiang, A.: Fault diagnosis of top drive gearbox based on the minimum entropy deconvolution resonance-based sparse signal decomposition. China Pet. Mach. **46**(18), 6–13 (2019)
12. Sun, X., Zhou, G., Yu, Y., et al.: Overview of prognostics and health management of mechanical equipment. Ordnance Ind. Autom. **35**(1), 30–33 (2016)
13. Shen, B., Chen, B., Zhao, C., et al.: Review on the research of deep learning in mechanical equipment fault prognostics and health management. Mach. Tool Hydraul. **49**(19), 162–171 (2021)
14. Zhang, H.: Research on a fault diagnosis system for drilling rig equipped with automatic pipe string processing system. China Pet. Chem. Stand. Qual. **40**(11), 52–54 (2021)

# Oriented Oilfield Structured Data Quality Assessment Model

Xue-song Su(✉), Wang Mei, Hui-fang Song, Jia Liu, and Shan Huang

Technology Inspection Center of Shengli Oilfield, Dongying, China
Xuesong.Su@outlook.com

**Abstract.** Since the utilization of data analytics in oil field industry, data mining has become increasingly important. Various decision-making algorithms derived from data are closely related to the quality of data, which makes data quality assessment an indispensable part of the intelligent construction of oilfield. General data quality assessment models are not suitable for centralized oilfield scenarios because the quality of datasets depends on their usage rather than a simple stacking of individual data units. For example, datasets containing data units with good quality yet serious homogeneity cannot meet the data requirements in deep learning. This paper is based on the theoretical model of process measurement and adopts the second-level fuzzy comprehensive evaluation model. We calculate the member-ship degree of each factor set based on the business demand by the AHP. The oriented oilfield structured data quality assessment model is then established. This model provides theoretical basis and technical support for oilfield data preprocessing, decision-making and staged evaluation of data governance.

**Keywords:** big data · data quality management · data quality assessment · information sampling · influence function

## 1 Introduction

With the widespread application of the Internet of Things and artificial intelligence technology in the oilfield industry, data is showing an exponential growth trend, and data quality issues have become increasingly important. Taking the intelligent inspection and

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20-22 September 2023.

This paper was selected for presentation by the IFEDC Committee following review of information contained in an abstract submitted by the author(s). Contents of the paper, as presented, have not been reviewed by the IFEDC Technical Team and are subject to correction by the author(s). The material does not necessarily reflect any position of the IFEDC Technical Committee its members. Papers presented at the Conference are subject to publication review by Professional Team of IFEDC Technical Committee. Electronic reproduction, distribution, or storage of any part of this paper for commercial purposes without the written consent of IFEDC Organizing Committee is prohibited. Permission to reproduce in print is restricted to an abstract of not more than 300 words; illustrations may not be copied. The abstract must contain conspicuous acknowledgment of IFEDC. Contact email: paper@ifedc.org.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 319–329, 2024.
https://doi.org/10.1007/978-981-97-0272-5_28

calibration of oilfield measuring instruments as an example, data quality issues such as noise, lack of labeling, and incompleteness during data acquisition, transmission, and processing can lead to serious decision-making errors in instrument drift and misalignment. Data quality assessment has become an essential part of oilfield intelligence construction. Existing data quality assessment models mainly focus on evaluating individual data units and do not consider the impact of the relationships between data units on data set quality. They also ignore the fact that different business requirements require evaluation of data quality dimensions. For example, a data set composed of high-quality pump operating data during normal operation cannot be considered a high-quality data set because it lacks data on pump leakage and other abnormal faults, making it impossible to perform effective machine learning. At the same time, in pump abnormal fault warning, data accuracy, completeness and other quality dimensions are mainly focused on, while in standard search and inspection, data timeliness, mobility and other quality dimensions are mainly focused on. This makes static data quality assessment models unable to deal with real oilfield data application scenarios. On the other hand, oilfields currently mainly focus on intelligent decision-making work based on structured data [1]. Therefore, how to scientifically, objectively, and realistically evaluate the quality of oilfield structured data has become an urgent problem to be solved.

Through a review of domestic and foreign literature, it was found that there is a lack of research on data quality assessment in oilfield data governance construction. In this study, the theoretical model of process measurement [2] was used as a reference, and the fuzzy comprehensive evaluation method [3] was adopted to comprehensively evaluate 20 data quality dimensions of oilfield structured data, such as completeness, consistency, accuracy, and timeliness, combined with the characteristics of oilfield structured data. The membership degrees of each factor set were calculated based on the specific business orientation through the Analytic Hierarchy Process [4]. Then, combined with the application of oilfield structured data in various intelligent decision-making projects [5–8], a directional oilfield structured data quality assessment model was finally constructed to provide a reference basis for the national oilfield information intelligence construction and data operation and maintenance management.

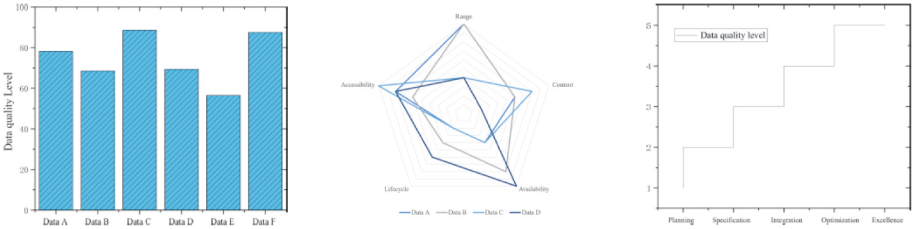
## 2 Methodology

### 2.1 Comprehensive Evaluation Model for Structured Data Quality in Oilfields

Developing a comprehensive data quality assessment model for oilfields not only provides a scientific and phased evaluation of the data governance process, but also identifies key areas for improvement based on quality defects identified in the evaluation model. Additionally, such a model provides guidance for data preprocessing in various machine/deep learning algorithms that rely on high-quality data.

**Selection of a Comprehensive Evaluation Model for Structured Data Quality**
Since 1990, various academic and industrial organizations and research institutions have conducted a series of studies on data quality, including its definition, problem characterization, and management applications. Starting in 2000, they began to establish their own data quality models and methodologies. Among the most representative are MIT's

Total Data Quality Management (TDQM) [9], which establishes a universal methodology for data quality based on the demands of industry and commerce and has been widely applied in various fields of data quality research. Another important data quality methodology is Data Warehouse Quality (DWQ) [10], which considers the diversity of data quality objectives and defines corresponding metadata. In addition, other scholars [11–13] have proposed various models for data usage, lineage, and other aspects. By summarizing and classifying various data quality evaluation models as well as other comprehensive evaluation models, we conclude that they can be mainly divided into three types: data scoring type, indicator distribution type, and stepwise type (See Fig. 1).



a. data scoring b. indicator distribution c. stepwise

**Fig. 1.** Types of data quality assessment models

The oilfield data governance is still in the development stage, and the data scoring type-based quality evaluation model cannot reflect the current stage of oilfield data governance. Meanwhile, the indicator distribution type-based model also has the drawback of being unable to provide a comprehensive evaluation. Therefore, this paper adopts a stepwise type evaluation model and, in response to the oilfield data quality evaluation facing the problem of the " sorites paradox," utilizes the method from fuzzy mathematics [3].

### Classification of Quality Levels for Structured Data in Oil Fields

The evaluation levels of the structured data quality assessment model for oilfields define the phased level of the data governance process. This article follows the standard of the Data Management Association (DAMA) and the Capability Maturity Model Integration (CMMI) in terms of data application [14]. Finally, the structured data quality levels for oilfields are divided into five levels: Level Five, Level Four, Level Three, Level Two, and Level One, in descending order. The higher-level evaluations must meet and include the advantageous characteristics of the lower-level evaluations and, to some extent, compensate for the defect structure of the lower-level evaluations. The specific meanings of each evaluation level are as follows:

– Level One (Initial Level): The data governance and planning process has begun with an IT infrastructure in place. Data quality activities are in a passive state with no formal data quality expectations, undefined data quality standards, no ability to measure data quality, and no ability to correlate data quality issues.
– Level Two (Standardized Level): Defective data can be repaired in a controllable manner by adding analysis process rules, and simple data errors can be identified and

warned. Specific expectations for data quality dimensions are defined and the ability to identify and measure data quality dimensions is in place.

- Level Three (Managed Level): A data management system and a defined process for data quality expectations are in place. Rules related to data quality are defined and can be used for data value validation, modeling, and interaction. Standards for the structure and format of all data elements are defined, including enterprise data specifications and metadata control measures.
- Level Four (Optimized Level): The compliance of data quality expectations and measured data quality dimensions can be associated with business impact analysis and interacted with. Data quality can be dynamically evaluated based on business objectives. The monitoring of data quality can be standardized and visualized, and self-repair can be performed.
- Level Five (Excellent Level): Suggestions for adjusting standards and processes for incremental data can be provided. Production data from business processes can be cross-validated, and the ability to define business-relevant data standards is in place. Rules for data quality services that can be followed are established. A continuous improvement strategy based on dashboard monitoring of the data lifecycle is fully implemented.

**Technical Requirements for Quality Levels of Oilfield Structured Data Quality Assessment Model**

The evaluation of structured data quality in oil fields at each level should meet the corresponding level's technical requirements for data quality (see Table 1), achieved through qualitative and quantitative analysis of data quality dimensions in data quality management. Quantitative measurement is entirely objective, using corresponding data dimension quality assessment calculation methods or tools for measurement. When qualitative measurement is required, data management personnel need to combine demand analysis, research data user requirements, and identify data quality issues or set new quality goals.

**Table 1.** Oilfield Structured Data Quality Evaluation Level and Data Quality Technology

| Quality Levels | Description of Data Quality Technical Requirements |
|---|---|
| Level 1 | No data quality assessment tool available |
| Level 2 | Uses fixed and static data quality assessment tools to obtain data standard rules |
| Level 3 | Introduces validation and assessment techniques based on business rules, with technical components for data authentication and reporting |
| Level 4 | Equipped with data visualization tools for data analysis, and can use defined business rules for data self-correction |
| Level 5 | Non-technical users can dynamically replace data quality dimensions and evaluation rules |

## 2.2 Comprehensive Evaluation Method for Structured Data Quality in Oil Fields

### Construction of Model Factor Set

*Establishment of Data Quality Dimensions*

In this model, the factor set is composed of data quality dimensions. To determine the data quality dimensions, we conducted a survey of 12 data quality assessment models including TDQM and DWQ from both domestic and international sources [9, 15–23] (see Table 2), and used UpsetPlot [24] to compare the dimensions across different models (see Fig. 2). We then selected 30 data quality dimensions that are in line with the business requirements of structured data in oil fields, based on the needs of the industry.

**Table 2.** List of Data Quality Evaluation Models

| Model Abbreviation | Model Full Name |
|---|---|
| AIMQ | A Methodology for Information Quality Assessment |
| AMEQ | Activity-based Measuring and Evaluating of product information Quality |
| CDQ | Comprehensive methodology for Data Quality management |
| CIHI | Canadian Institute for Health Information |
| DQA | Data Quality Assessment |
| DWQ | Data Warehouse Quality |
| DaQuinCIS | Data Quality in Cooperative Information Systems |
| IQM | Information Quality Measurement |
| ISTAT | Italian National Bureau of Census |
| QAFD | Methodology for the Quality Assessment of Financial Data |
| TDQM | MIT's total Data Quality Management |
| TIQM | Total Information Quality Management |

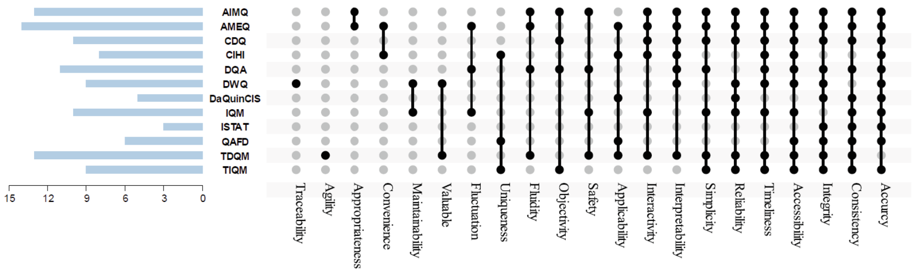


**Fig. 2.** Cross-comparison of Data Quality Dimensions across Data Models

*Factor Set Partitioning*

1. Constructing the factor set $U = \{u_1, u_2, \cdots, u_n\}$ with the obtained data quality dimensions as mentioned above.

2. Partition the factor set $U$ into several groups $U = \{u_1, u_2, \cdots, u_k\}$, where

$$U = \bigcup\nolimits_{i=1}^{k} U_i \wedge \left(U_i \cap U_j = \phi(i \neq j)\right) \tag{1}$$

3. We label $U$ as the first-level factor set, and $U_i$ as the second-level factor set. Finally, the 30 data quality dimensions are hierarchically clustered and divided into 20 first-level factor sets.

*Methods for Calculating Quality Dimension Indicators*

In evaluating the quality dimensions of data, differences in methods and processes can lead to significantly different results. To address this issue, we propose a formalized calculation process for each data quality indicator. For example, Dr. Han Liu [25] proposed a formula for calculating data completeness as follows:

$$u_{21} = \frac{1}{T}\sum\nolimits_{i=1}^{T} \frac{R_{ij}}{\bigcup_{j=1}^{S} R_{ij}} \tag{2}$$

where $u_{21}$ represents data completeness, $S$ is the total number of data sources, $T$ is the number of queries, and $R_{ij}$ is the set of results returned by the $j$-th data source after the $i$-th query. Equation (2) indicates that after sampling or discretizing the required data sources, the integrity of the data can be determined by querying and sampling the results returned by multiple data sources.

However, in actual evaluation of structured data quality in oilfields, data quality still ranges from initial level to management level. Using the same type of calculation method as Eq. (2) not only takes a lot of time but also yields results that are not significantly different from traditional calculation methods (non-null values/total values), which does not contribute to breakthrough improvement in the final quality evaluation. Furthermore, the calculation method [25] faces the problem of dimension explosion in time complexity when dealing with massive objects, elements, and feature values in oilfield PCS (Production Control System) data sources. To address this, our proposed model simplifies the calculation methods for each data quality evaluation to better adapt to the characteristics of current structured data in oilfields.

*Establishment of a Comment Bank for Evaluation*

Establishment of a comment bank for evaluation based on the rating determination and description in Sect. 2.1:

$$V = \{\text{Initial}v_1, \text{Standardized}v_2, \text{Managed}v_3, \text{Optimized}v_4, \text{Excellent}v_5\} \tag{3}$$

*Determination of Weightings for Each Factor*

The comprehensive evaluation of data quality not only relates to the assessments of various factors, but also depends on the weightings between these factors. Methods for determining weightings include the entropy weight method, Delphi method, weighted average method, among others. However, weightings obtained from methods such as the entropy weight method that are based on the data itself cannot be dynamically adjusted according to changing business needs. Therefore, this paper favors the use of the Analytical Hierarchy Process (AHP) as a basis for efficient handling of the interrelationships between multiple objectives to obtain weightings. In the specific implementation process, due to the limitations of AHP with regard to the consistency indicator RI, we adopt a Top 10 factors set (the 10 factors that have the greatest impact on specific business needs), and then use the AHP method to dynamically establish weightings for each factor:

$$A = [a_1, a_2, \cdots, a_{10}] \tag{4}$$

where $a_i$ represents the weighting for the $i$-th factor, and satisfies

$$\sum\nolimits_{i=1}^{10} a_i = 1 \tag{5}$$

*Construction of Fuzzy Comprehensive Evaluation Matrix*

For each indicator $u_i$ in each factor set, its affiliation score in each comment bank is a fuzzy subset on $V$. The evaluation of indicator $u_i$ is denoted as:

$$R_i = [r_{i1}, r_{i2}, r_{i3}, r_{i4}, r_{i5}] \tag{6}$$

The fuzzy comprehensive evaluation matrix for each indicator represents a fuzzy relation matrix from $U$ to $V$, denoted as:

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{15} \\ r_{21} & r_{22} & \cdots & r_{25} \\ \vdots & \vdots & \ddots & \vdots \\ r_{101} & r_{102} & \cdots & r_{105} \end{bmatrix} \tag{7}$$

*Performing Comprehensive Evaluation of Data Quality*

1. If there is a fuzzy relation $R = \left(r_{ij}\right)_{10\times 5}$ from $U$ to $V$, a fuzzy transformation can be obtained from $R$ as follows:

$$T_R : F(U) \rightarrow F(V) \tag{8}$$

2. Using the transformation above, the comprehensive evaluation result of data quality can be calculated as:

$$B = A \cdot R \tag{9}$$

3. The integrated evaluation can be regarded as a fuzzy vector on $V$, denoted as:

$$B = [b_1, b_2, \cdots, b_5] \tag{10}$$

4. If $max[b_1, b_2, \cdots, b_5] = b_k$, the data quality evaluation will be classified as $k$.

# 3 Method Evaluation

## 3.1 Engineering Project Background

Since 2019, we have been focusing on intelligent analysis of spare parts consumption patterns in oilfield dynamic equipment operation and maintenance, which not only helps to predict the life cycle of key vulnerable parts of oilfield equipment, but also provides scientific advice from procurement and maintenance perspectives. For data sources, we extracted material distribution data from two different regions of the material management system, as well as well daily report, well real-time data table, pump real-time data table from the PCS management system, and switch record, pump maintenance manual collection data from the EPBP management system. In this paper, we optimized the main control indicators for the analysis of dynamic equipment spare parts consumption patterns, determined the priority weights of data quality dimensions, preprocessed and modeled the data, and finally obtained a full life cycle prediction model.

After the research results were implemented, we sorted out the data quality in the analysis of equipment consumption patterns, evaluated the data quality on different regions and time dimensions, and conducted correlation analysis based on the corresponding data and the effect of full life cycle prediction.

## 3.2 Results of Data Quality Assessment

We conducted a oriented data quality assessment on the data collected from June 2019 to June 2022, and found that only the data from region 1 during June 2021 to June 2021 achieved Level 3 in the management level, while the data quality evaluation of other regions remained at Level 2 in the standard level. We used affiliation score to represent their quality level within the evaluation period. On the other hand, by sampling the data distribution estimated for the equipment's full life cycle, we found that it followed an average distribution. Therefore, we used $F_1\_score$ as the evaluation metric to assess the performance of our example model. The cross-comparison results were obtained as shown in Fig. 3.

The calculation formula for $F_1\_score$ is as follows:

$$F_1\_score = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

## 3.3 Engineering Project Analysis

1. Through comparing the $F_1\_score$ of two regions, the results show that the overall engineering performance of region 2 is lower than that of region 1. By comparing the

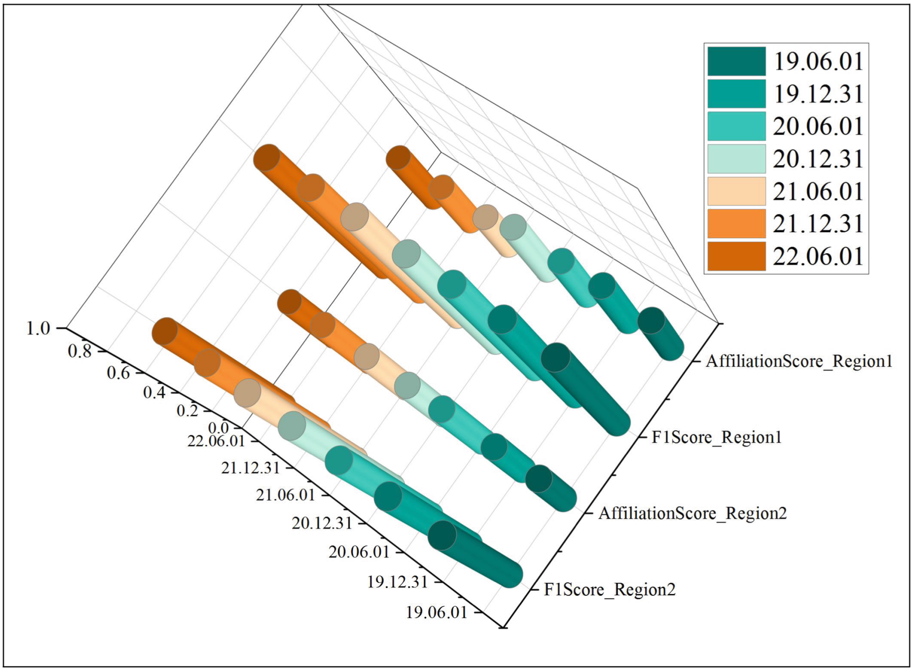


**Fig. 3.** Relationship Diagram between Data Quality Assessment and Engineering Examples' Effectiveness

corresponding data quality assessment scores, the results show that there is a positive correlation between data quality and model application performance. On the one hand, it illustrates that data quality influences the model performance of intelligent decision-making. On the other hand, it also reflects the deficiency of the generalization ability in our life cycle prediction model application, which is specifically manifested as the model's over-reliance on local data quality and inability to associate and predict across regions.

2. Through field visits, it was found that region 1 has a better management system than region 2, specifically reflected in the compliance recording of material distribution management and pump maintenance data. Therefore, in terms of quality dimensions such as data completeness, region 1 is superior to region 2, which ultimately manifests in the data quality comprehensive evaluation, where region 1 is higher than region 2, fully reflecting the important role of data management system in data governance.
3. By comparing the data from region 1 starting from June 2021, the analysis shows that the $F_1\_score$ has increased by about "5%", and its data quality assessment has also risen from level two (specification level) to level three (management level), demonstrating the effectiveness of data governance in region 1 over the past year.
4. By comparing the data quality of region 2 in the past three years, the results show that although the data governance process is slowly progressing, there is a slight improvement in model performance. This illustrates the iterative optimization capability of

the life cycle prediction model and also indirectly indicates that data pre-processing methods can to some extent make up for the deficiencies in data quality.

## 4 Conclusion

In this paper, we analyzed and compared mainstream data quality evaluation models at home and abroad, and developed a structured data quality evaluation method for oilfields, including various quality dimensions and metrics. We provided scientific evidence and theoretical support for oriented evaluation of structured oilfield data. However, for unstructured data in oilfields, such as speech, text, images, and videos, the quantitative models and evaluation methods are significantly different, as text data needs to consider contextual relevance and expression quality. As oilfields increasingly adopt intelligent construction towards unstructured data, future research should analyze and study semi-structured/unstructured data in oilfields to provide comprehensive and effective guidance for oilfield data quality evaluation.

**Acknowledgments.** The research is supported by Sinopec Shengli Oilfield Intelligent Engine for Oilfield Safety and Environmental Protection project under grant number YKB2302.

## References

1. Liang, W.F.: Constructed achievements and prospects of the intelligent application system for the oilfield development. Daqing Petrol. Geol. Dev. **38**(5), 283–289 (2019)
2. Liu, S.F., Gao, W.X., Xu, W.Y.: Research and improvement of a software process measurement model. Microcomput. Inform. **26**(33), 17–19 (2010)
3. La, Z.: A Fuzzy-set-theoretic approach to the compositionality of meaning: propositions, dispositions and canonical forms. J. Seman. **2**(3–4) (1983)
4. Saaty, R.W.: The analytic hierarchy process—what it is and how it is used. Mathem. Model. **9**(3), 161–176 (1987)
5. Chen, S.H.: Research on application of image enhancement and recognition technology based on deep learning in oilfield work area. University of Electronic Science and Technology of China (2021)
6. Wang, Q., Li, A.R., Ren, J.W., et al.: Exploration of deep learning technology in remaining oil prediction of Y oilfield. Surv. Mapping Geol. **4**(3), 15–18 (2022)
7. Wang, H.L., Mu, L.X., Shi, F.G., et al.: Production prediction at ultra-high water cut stage via recurrent neural network. Petrol. Explor. Dev. 2020107 (2020)
8. Zhong, Y.H., Wang, S.N., Luo, L., et al.: Knowledge mining for oilfield development index prediction model using deep learning. J. Southwest Petrol. Univ. **42**(6), 63 (2020)
9. Wang, R.Y.: A product perspective on total data quality management. Commun. ACM **41**(2), 58–65 (1998)
10. Laws, R., Gillespie, S., Puro, J., et al.: The community health applied research network (CHARN) data warehouse: a resource for patient-centered outcomes research and quality improvement in underserved, safety net populations. EGEMS **2**(3) (2014)
11. Fang, Y.L., Yang, D.Q., Tang, S.W., et al.: Data quality managements in data warehouse. Comput. Eng. Appl. **39**(13), 1–4 (2003)
12. Fang, Y.L., Yang, D.Q., Tang, S.W., et al.: Data schedule serialization in data transformations. Comput. Eng. Appl. **39**(17), 4–6 (2003)

13. Liu, Z.H., Zhang, Q.L.: Research overview of big data technology. J. Zhejiang Univ. **48**(6), 957–972 (2014)
14. Constantinescu, R., Iacob, I.M.: Capability maturity model integration. J. Appl. Quantit. Methods **2**(1), 31–37 (2007)
15. Batini, C., Cabitza, F., Cappiello, C., et al.: A comprehensive data quality methodology for web and structured data. In: Proceedings of the 2006 1st International Conference on Digital Information Management. IEEE (2006)
16. Boretti, E.: AIB-ISTAT statistics: the first time for Italian public libraries. Perform. Measur. Metrics **6**(1), 32–38 (2005)
17. Joseph, K., Mmath, J.F.: Validation of perinatal data in the Discharge Abstract Database of the Canadian Institute for Health Information. Chronic Diseases Injur. Canada **29**(3) (2009)
18. Lee, Y.W., Strong, D.M., Kahn, B.K., et al.: AIMQ: a methodology for information quality assessment. Inform. Manag. Commun. ACM **40**(2), 133–146 (2002)
19. Marchetti, C., Mecella, M., Scannapieco, M., et al.: Data Quality in Cooperative Information Systems. Encyclopedia of Data Warehousing and Mining, pp. 297–301. IGI Global (2005)
20. Pipino, L.L., Lee, Y.W., Wang, R.Y.: Data quality assessment. Commun. ACM. **45**(4), 211–218 (2002)
21. Sitawati, H.D., Ruldeviyani, Y., Hidayanto, A.N., et al.: Data quality improvement: case study financial regulatory authority reporting. In: proceedings of the 2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE). IEEE (2022)
22. Su, Z., Jin, Z.: A methodology for information quality assessment in the designing and manufacturing processes of mechanical products. Information Quality Management: Theory and Applications, pp. 190–220. IGI Global (2007)
23. Woodall, P., Parlikad, A.K., Lebrun, L.: Approaches to information quality management: State of the practice of uk asset-intensive organisations. Asset Condition Informat. Syst. Decision Models, 1–18 (2012)
24. Chen, C., Chen, H., Zhang, Y., et al.: TBtools: an integrative toolkit developed for interactive analyses of big biological data. Mol. Plant **13**(8), 1194–1202 (2020)
25. Liu, H.: The research on key issues of data quality management, assessment and detection in big data environment. Jilin University (2019)

# Research on Automatic Classification of Premium Threaded Connections Make-Up Torque Curve Based on CNNs with Data Augmentation

Zi-han Ma(✉), Yu Fan, We Luo, Chuan-lei Wang, Lang Zhou, Du Wang, and Yun-qi Duan

Engineering Technology Research Institute of Petrochina Southwest Oil and Gas Field Company, Chengdu, Sichuan, China
447905923@qq.com

**Abstract.** Leakage of premium thread connections tubing is the main reason for annulus pressure and affecting well integrity level. At present, helium gas seal detection and manual monitoring of make-up torque curve are mainly used to ensure the integrity of gas seal of well string. However, the helium seal detection environment is static detection, which fails to describe the air tightness of the tubing under complex downhole load; The manual monitoring of the make-up torque curve depends on the field engineer with certain experience to check the standard curve one by one. The results are greatly affected by subjective factors, so it is difficult to unify the measurement standard. Therefore, a machine learning method based on convolutional neural network (CNN) is proposed to automatically identify and classify the makeup torque curve of special threaded tubing. In order to achieve this goal, firstly, according to the failure of gas seal detection, manufacturer's manual and field experience, the categories of makeup torque curve are divided, including typical curve, acceptable curve and unacceptable curve. Secondly, in order to improve the model training accuracy and further improve the prediction results, the data expansion technology is used to expand the training database. Finally, the multilayer convolutional neural network model is built and

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20-22 September 2023.
This paper was selected for presentation by the IFEDC Committee following review of information contained in an abstract submitted by the author(s). Contents of the paper, as presented, have not been reviewed by the IFEDC Technical Team and are subject to correction by the author(s). The material does not necessarily reflect any position of the IFEDC Technical Committee its members. Papers presented at the Conference are subject to publication review by Professional Team of IFEDC Technical Committee. Electronic reproduction, distribution, or storage of any part of this paper for commercial purposes without the written consent of IFEDC Organizing Committee is prohibited. Permission to reproduce in print is restricted to an abstract of not more than 300 words; illustrations may not be copied. The abstract must contain conspicuous acknowledgment of IFEDC.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 330–345, 2024.
https://doi.org/10.1007/978-981-97-0272-5_29

trained and verified based on the data. In the model verification stage, four comprehensive evaluation methods are used: typical rate (TR), acceptable rate (AR), unacceptable rate (UR) and accuracy (P). The proposed CNNs model is evaluated accurately and compared with state-of-art machine learning algorithms such as SVM and logistic regression.

**Keywords:** Sustained casing pressure · Premium thread · Machine learning · Convolutional neural networks · Data augmentation

## 1 Introduction

With the progress in the development of HTHP (high temperature, high pressure) sour gasfield in recent years, string screw leakage, failures of packers and abnormal pressurization of annulus are increasingly exposed. These problems may directly affect the successful operation of gas wells and the production safety in later stages [1–10]. With these concerns, the HTHP sour condition puts forward high requirements for completion tools and strings, especially the loads placed on ordinary connections with certain well conditions exceed those of an API class-1 qualified connection. Therefore, for the sake of security, a growing number of natural gas company initiated the utilization of premium threaded connections (PTCs) and permanent packer as the completion strings which are designed to withstand today's toughest well conditions. Even so, a majority number of oilfield cases has confirmed that seal failure still happened on the PTCs frequently, which eventually developed into sustained casing pressure(SCP) and resulted in a large number of economic losses. Through finite element analysis it has been found that the sealing performance of the joint depends on the sealing surface, the shoulder and the contact stress on the thread [11]. Torque versus turn make-up graphs are useful to evaluate if a joint is correctly made-up. Hereby, manufactures released their own correct "torque signature" to ensure threads connected in appropriate stress. In order to achieve this purpose, there must be a system of ensuring that a competent or experienced person is on the drill floor at all times to check any make-up graphs. However, the evaluation standards of each engineer are not uniform, and there are errors and uncertainties in manual monitoring. Hence, it is inevitable to introduce artificial intelligence to monitor the PTCS make-up process.

In recent years, artificial intelligence technologies flourish in various fields and manifest immense potential in the industry toward smart wells and intelligent fields. Especially the artificial neural network, fuzzy logic, and evolutionary algorithms are common among AI techniques being applied today in oil and gas reservoir simulation, production and drilling optimization, drilling automation and process control, and data mining [12]. In 2002, Jose Finol [13] discussed the fuzzy logic method which is used to address the inverse problem of permeability prediction from NMR data. AI-Fattach S.M presented an artificial neural network method to predict the natural-gas production in the unite-states for the next 15 years [14, 15]. Vikas Bhushan & Simon Christopher Hopkinson developed an artificial intelligence technique called case-based reasoning to identify reservoir analogues systematically with sparse information [16]. However, a variety of integrated forms of artificial intelligence from neural networks to fuzzy logic, have made solid steps toward becoming more prevalent in the mainstream of the oil and gas industry. The AI related to image classification and visual recognition still waits to be explored in depth.

In decades development of machine learning there is a general consensus that deep learning has achieved excellent performance in various recognition problems. Among different types of deep neural networks, the research on convolutional neural networks (CNNs) has been most extensively made. Benefit from the rapid growth in amount of data and the great advancement in image process techniques, the CNNs has thrived and made remarkable breakthroughs on target detection [17–19], pattern recognition [20, 21] and image classification problems [22–24]. Compared with traditional image processing methods [25], the CNNs can extract image features efficiently, and better complete classification and recognition tasks.

Although the CNNs has made great successfulness in different fields such as biological image classification [18], identification of defects in the centrifugal pump [26] and classification of the high-resolution imagery scene [27], the gaps of application in oil and gas industry are still waited to be filled. More recently, a few researches have been made to explore the application of CNNs in image classification on sucker rod pumping wells. Yi peng took advantage of deep convolutional neural networks to analyze the relationships between the electrical data and corresponding dynamometer card in different conditions. The results are used to diagnose the suck-rod pumping wells [28]. SA Sharaf developed a machine learning algorithm to achieve pump card classification, where the pump card is a plot of load versus position on the pump's plunger. The machine-learning techniques can simulate the human visual-interpretation process to identify and diagnose the pump operating conditions [29]. With the help of CNNs, the labor time consuming on image classification works has been significantly alleviated as well as the requirement of deep expertise in the domain has been drastically reduced.

Inspired by the achievements of CNNs in image classification works, we presented a CNNs model with data augmentation technique to classify the make-up torque graphs of PTCs. The training datasets are labeled as typical curves, acceptable curves and unacceptable curves in accordance with expert domain knowledge. In order to improve the robustness and accuracy of the proposed model, the data augmentation technique has been introduced to enlarge the training datasets. Moreover, to better evaluate the classification performance of the model, four measures, namely Typical rate (TR), acceptable rate (AR), unacceptable rate (UR) and Precision are used in the proposed model. The experimental results are compared with state-of-art methods to demonstrate the feasibility of our model.

The rest of this article is structured as follows. Section 2 includes the detail of datasets, and then the proposed data augmentation and image processing techniques are described. The Sect. 3 illustrates the experimental results and the discussion of proposed method.

## 2 Materials and Proposed Model

### 2.1 Materials

Premium threaded connections (PTCs) of tubing are an optimal selection of pipe screws in the MX Gas Field of the Sichuan Basin. The field data are make-up graphs of PTCs provided by MX gas well in one block in Sichuan Basin. Three types of curve pattern could be concluded in accordance with filed domain expert experience and Running_Manual of PTCs manufacturers, which are typical curve, acceptable curve and unacceptable curve. The samples of these datasets are illustrated in Fig. 1. These graphs form the datasets used to train the CNNs model in make-up graphs classification and verify the effectiveness of the proposed CNN with data augmentation method.

Typical curve of PTCs make-up torque is composed of thread &seal interference section, shouldering section and release section. In thread & seal section, the threads

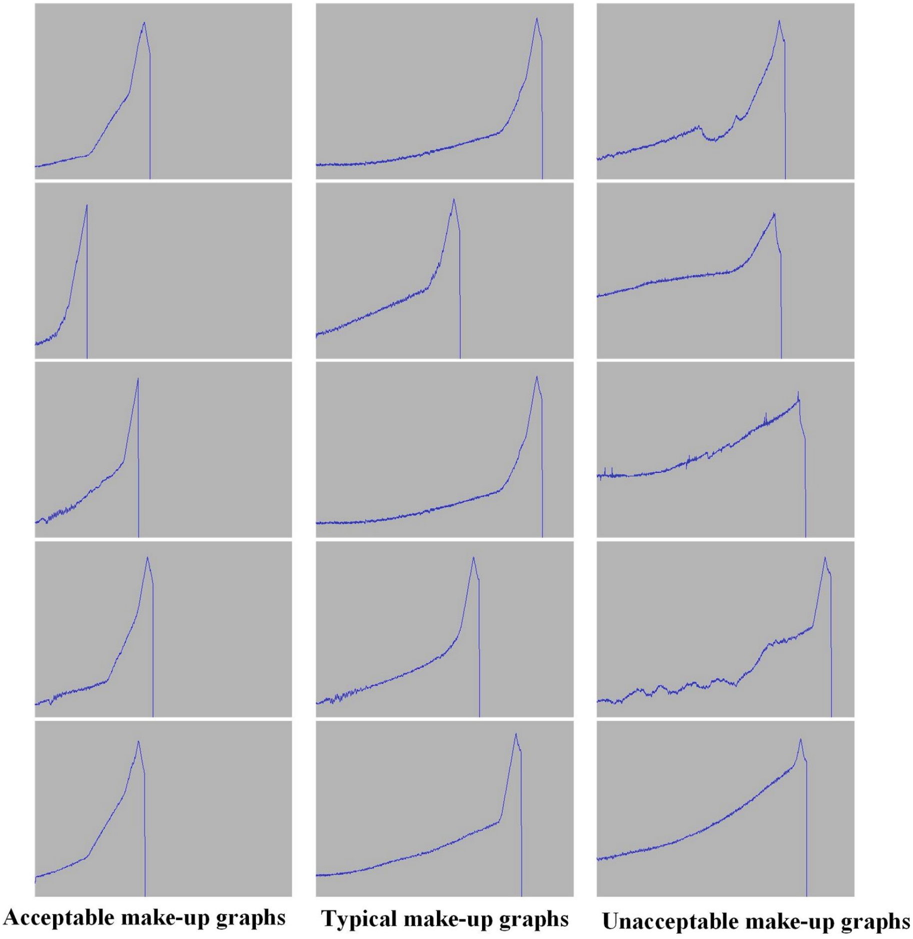


**Fig. 1.** Samples of datasets from MX gas well

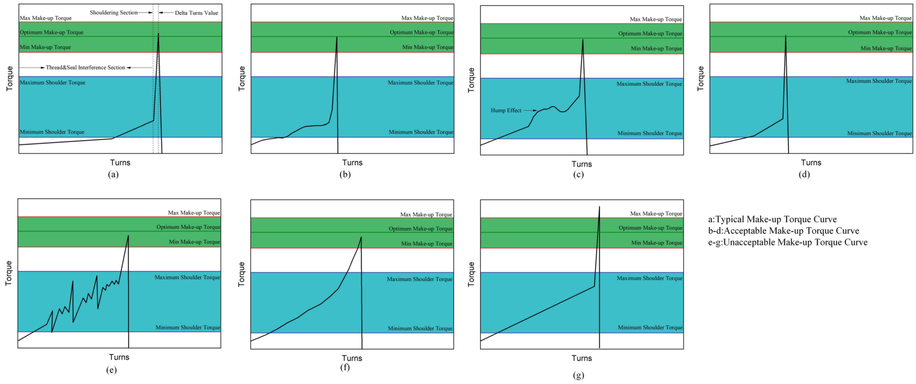


**Fig. 2.** Examples of different make-up torque graphs

start to interfere and the torque increase at early time is relatively low in recording turns. Then, it followed by shouldering section where the shoulders of the connections meet, and the sealing function starts to work when the torque increases sharply. This characteristic leads us to find the difference between typical curve and the others. Finally, a sharp descending of torque represents the release section. Figure 2 shows the example of the typical curve.

Acceptable curves meet the seal and connection requirements where curve profile is not exactly the same as the typical curve. Such as the slight oscillations are noted when threads are interfered at early time. Meanwhile, a few curves show the hump effect in thread & seal section which are still allowed. There are two possible reasons. One is that dirt between threads. The other is running compound excess. The examples of these curves are shown in Fig. 2 and the particular description of each curve is illustrated in Table 1.

Although running casing or tubing is a relatively low-risk activity and PTCs products have been designed with ease and reliability of make-up as priorities. There still remains a small chance of incorrect make-up which is classified as unacceptable curve. If non-defined torque shoulder took place, the make-up torque would gradually increase during thread & seal and shouldering section which is defined as unacceptable curve. Another typical pattern of unacceptable curve is erratic fluctuation of torque during thread & seal section. Figure 2 represents samples for unacceptable curves and the detailed explanation is described in Table 1.

**Table 1.** Detailed explanation of make-up torque graphs

| Number | Description | Possible Causes | Consequence |
|---|---|---|---|
| a | Typical curve | – | – |
| b | acceptable curve | High friction running compound | – |
| c | acceptable curve | Running compound excess dirt between threads | – |
| d | acceptable curve | – | – |
| e | unacceptable curve | Over torque high momentum | over load thread galling |
| f | unacceptable curve | cross threads misalignment damaged threads | Leak thread galling |
| g | unacceptable curve | galled threads threads with razor edges | thread galling leak |

## 2.2 Proposed Model

### 2.2.1 Data Augmentation

It is the common knowledge that the more data a machine learning algorithm access to, the better classification performance it achieves. Data augmentation is an effective way to increase the amount of training data by using effective transformations in the original datasets. With the help of data augmentation, the proposed method achieves better accuracy, particularly true in this project involved expert domain knowledge where collecting sufficient data is heavily hindered. Besides, the meagre amount of data also results in problems of overfitting. Hence, in order for the model to be successful, large amount of data is required to train and validate the machine learning algorithm. The most effective data augmentation technique should be learnt and implemented.

At present, various data augmentation techniques have been applied in specific situation. The main techniques fall under the category of data warping [30], which is an approach which seeks to directly enlarge the input data in data space. The original patterns of images are transformed by rotations, flips, contrast variations and translations. For instance, vertical flips, the pixel position at coordinate [x, y] in input image is altered at coordinate [x, −y] in new image. In this case, for a dataset of size N, the final dataset would be increased to 2N with the of help of proposed method. Considering all the transformations, a common form expresses affine transformation of the original image.

$$y = Wx + b \tag{1}$$

where the *x, y* represents pixel coordinates. The value of *w* and *b* depends on the augmentation technique we choose.

### 2.2.2 Convolutional Neural Networks Architecture

In this section we go into detail of convolutional neural networks (CNNs). It is formed from three types of layers, namely convolutional, pooling, fully connection layers. The architecture of proposed model is shown in Fig. 3. For image classification problem, the inputs are RGB images which are three-dimensional matrix represented as x ∈ $\mathrm{R}^{\mathrm{h \times w \times c}}$ (h: height, w: width, c: channel).

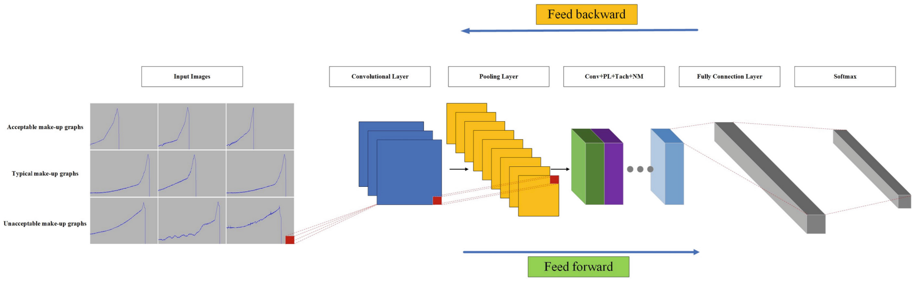


**Fig. 3.** Architecture of convolutional networks

Firstly, the function of convolutional layer is to learn the features of the input data. Each convolution layer consists of multiple convolutional kernels which are used to take convolution operation on inputs to receive different feature maps. Each neuron of a feature map corresponds to a region of adjacent neurons in preceding layer. The complete feature maps are calculated by using a bank of kernels and adding a bias which is shown as Eq. 2.

$$z_{i,j,l} = \sum_{i=1}^{h} \sum_{j=1}^{w} \sum_{d=1}^{c} w_l^{c^T} x_{i,j}^c + b_l^c \tag{2}$$

where i, j represents the input image pixel coordinate, and the c is the number of channels. l stands for the l-th feature map. $\mathrm{w}_\mathrm{l}^{\mathrm{c}^\mathrm{T}}$ and $\mathrm{b}_\mathrm{l}^\mathrm{c}$ are the weight and bias of the l-th filter of c-th channel respectively. Specifically, the size of feature map is determined by the size of kernel and stride we selected. Mathematically, the dimension of l-th feature maps, D, turns out to be governed by the following formula:

$$D = \frac{n+2p-f}{s} + 1 \tag{3}$$

where n is the input image size, n × n. f is the filter size, f × f. s and p is the stride and padding we choose respectively. It is worth to mention that we take D round down to the nearest integer when fraction is not an integer.

Secondly, in order to solve the problem of insufficient expression of linear model, the activation function has been applied to introduce nonlinearities to CNN and endows the network with the ability to detect nonlinear features. Let a(z) express the nonlinear activation function. The mathematical expression is described as formula 4.

$$a_{i,j,c}^l = a\left(z_{i,j,c}^l\right) \tag{4}$$

There exists various types of activation functions, out of which the typical functions are Sigmoid, Tanh, and ReLU [31, 32]. In order to accelerate the convergence in each layer, the Tanh activation function has been chosen in this article. The equation and figure of tach function is shown as follow (Fig. 4).

$$a(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{5}$$

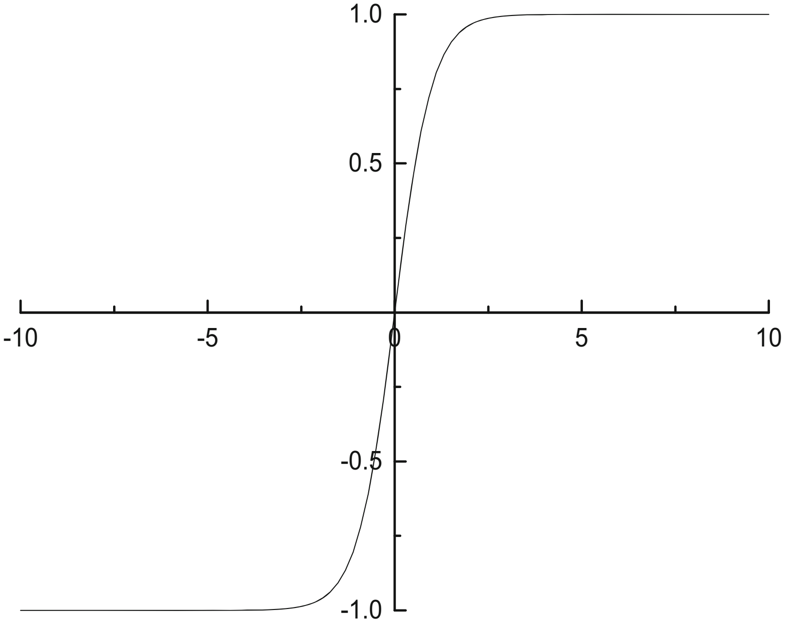


**Fig. 4.** The graph of Tanh function

The pooling layer is used to reduce the size of feature maps to speed up computation as well as to improve its robustness in classification. Out of which the typical pooling operation is maximum pooling. It is normally located between two convolutional layers The pooling operation is similar to convolutional operation. These layers have kernels of small size, which cover the entire image through shifting. Instead of taking dot product between input and kernel, maximum cooling operation returns the maximum value for each sub-region.

After several convolution layers and pooling layers, there may be one or more fully-connected layers which aim to perform high-level reasoning [33–35]. The fully connection layer flats the previous feature maps into 1-dimensional vector and integrates features to specific value. Note that fully connection layer is normally over two layers as it has better performance in solving nonlinear problems.

The last layer of CNN would be the output layer. For classification purpose, the Softmax operator is commonly used [36]. The Softmax is used to calculate the probability distribution of each class over all setting classes. Mathematically, operation of Softmax layer is shown as:

$$P_I = \frac{e^{z^I}}{\sum_j e^{z^J}} \tag{6}$$

where the $z^I$ stands for the inputs of Softmax, and the J is the number of classes. With the help of Softmax operator, the output value of each class is transferred into probability distribution in the range of (0, 1), and the sum of all the classes probability turns to 1. The cross entropy loss function can be used to evaluate the similarity between the predicted and actual output, the formula is shown as follow:

$$\tau = \sum_{J}^{I} y_J \log p_I \tag{7}$$

where $y_J$ represents the target outputs, $p_I$ denotes the predicted outputs. For the classification problem, the loss function indicates the error between actual outputs and predicted outputs. The training process of CNNs can be regarded as a problem of global optimization, which is to find the best set of parameters to minimize the error. We choose stochastic gradient descent method (SGDM) to minimize the loss function. The mathematical formula is shown as:

$$\theta_{l+1} = \theta_l - \alpha * \nabla E(\theta_l) \tag{8}$$

where l is the iteration number, alpha is the learning rate, θ is the loss function. This SGDM takes small steps at each iteration in the direction of the negative gradient of the loss to minimize the error and update the parameter vector.

# 3 Experiments and Analysis

## 3.1 Implementation Details

In this section, the details of the proposed method are explained. The datasets, the architect of CNNs model is introduced in detail, and followed by three measures which are presented to assess the performance of the classification models. Then, the results of our experiments and the comparison of state-of-art methods are discussed exhaustively.

For the purpose of extracting specific features of the make-up torque graphs, CNNs model with data augmentation techniques has been used. This model consists of convolution layers, pooling layers, batch normalization layers and fully-connection layer. Specifically, the architect and the hyperparameters of the proposed CNNs classification model are shown in Table 2 and Table 3 respectively. All frameworks are implemented using Intel(R) Core(TM) I7-9700 CPU @3.00 GHz @8 GB RAM.

In the experiments, in order to evaluate the performance of the proposed method, these measures were calculated: typical rate(TR), acceptable rate(AR), unacceptable rate(UR), precision [37]. The description of these measures is displayed as follows:

- Typical rate: the ratio of typical curves which are classified correctly
- Acceptable rate: the ratio of acceptable curves which are classified correctly
- Unacceptable rate: the ratio of unacceptable curves which are classified correctly
- Precision: manifest the accuracy of the proposed method to all the instances in the dataset

$$TR = \frac{NT}{NT+FNT} \tag{9}$$

**Table 2.** Hyperparameters of proposed method

| Hyperparameters | Value |
|---|---|
| Learning rate 1 | 0.01 |
| Epochs | 10 |
| Activation function | Tach |
| Loss function | Stochastic Gradient Descent |
| Execution environment | Cpu |
| input image size | 500 × 500 × 1 |

**Table 3.** Architect of CNNs

| Layer | Layer name | Layer size |
|---|---|---|
| 1 | image input layer | 500 × 500 × 1 |
| 2 | convolution 1 | 10 × 10 × 8 with stride 3 |
| 3 | batch normalization layer | – |
| 4 | Tach activation layer | – |
| 5 | Max pooling layer | 5 × 5 with stride 2 |
| 6 | convolution 2 | 8 × 8 × 16 with stride 3 |
| 7 | batch normalization layer 2 | – |
| 8 | Tach activation layer 2 | – |
| 9 | max pooling layer 2 | 3 × 3 with stride 2 |
| 11 | convolution layer 3 | 3 × 3 × 16 with stride 1 |
| 12 | batch normalization layer 3 | – |
| 13 | Tach activation layer 3 | – |
| 14 | fully connected layer | 3 |
| 15 | softmax | – |
| 16 | classification | – |

$$AR = \frac{NA}{NA+FNA}$$
$$UR = \frac{NU}{FNU + NU} \tag{10}$$

$$precision = \frac{NT+NA+NU}{S} \tag{11}$$

where the NT means the number of typical curves which are classified correctly, the NA is the number of acceptable curves which are classified correctly, and the NU stands for the number of unacceptable curves which are classified correctly. On the contrary, FNT, FNA, FNU represents the number of corresponding class which is classified incorrectly. Moreover, S is the total number of instances.

## 3.2 Experimental Results and Comparison

In this section, the performance of the proposed CNN-with data augmentation method for make-up torch graph classification is evaluated. A series of experiments on the measures (TR, AR, UR, Precision) of CNNs model are conducted in same datasets and environment. Each dataset consists of 1000 graphs where 800 graphs are randomly chosen as the training set, the other is set as validation set. The results are presented as comparison tables and line graphs. Moreover, the comparison between our method and state-of-art image classification methods is implemented. The results are used to verify the robustness of our classification model.

### 3.2.1 Training and Validation of Proposed Method

To evaluate the classification performance of the proposed method with data augmentation, we carry out our experiments in accordance with the work flow shown in Fig. 2. Firstly, the augmentation net implements random rotation and axis-translation on each input images and returns augmented images of size 500 × 500 × 1. Afterward, these images are inputted into the CNNs net to train the model, where the architect of CNNs model is shown in Table 2 and the training hyperparameters are shown in Table 3. The results of the experiments are illustrated below:

**Table 4.** Measures of proposed method

| Measures | Accuracy |
|---|---|
| TR | 0.9830 |
| AR | 0.9920 |
| UR | 0.9918 |
| precision | 0.9883 |

As can be seen from the Table 4, the accuracy of TR, AR, UR and Precision reaches 98.30%, 99.20%, 99.18%, 98.83% for typical curves, acceptable curves and unacceptable curves datasets respectively. This high accuracy confirms that the CNNs with data augmentation performs well on classification work. Moreover, detailed evaluation methods verify the performance of the proposed model on different data sets. Figure 5 indicates the trend of training and validation accuracy while running the experiments, which further demonstrates the feasibility and robustness of proposed model.

### 3.2.2 Effect of Augmentation Method (with/without)

To assess the effectiveness of data augmentation technique, a comparison between the proposed CNN model with and without data augmentation is presented in Table 5. The comparative experiments are conducted under the same conditions and original datasets.

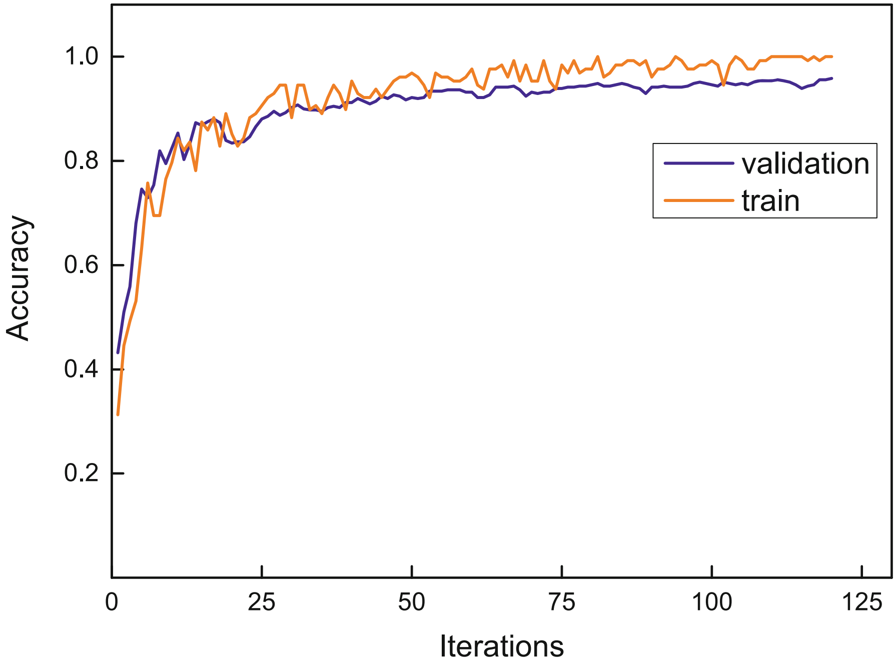


**Fig. 5.** Plot for accuracy on training and validation of proposed model

**Table 5.** Comparison of proposed CNNs model with and without data augmentation

| Dataset | Accuracy | |
|---|---|---|
| | With data augmentation | Without data augmentation |
| AR | **0.9830** | 0.9059 |
| TR | **0.9920** | 0.9295 |
| UR | **0.9918** | 0.9265 |
| Precision | **0.9883** | 0.9235 |

AS can be seen from Table 5, results show improvement in accuracy by 7.71%, 6.25%, 6.53% and 6.48% on AR, TR, UR and precision for typical curves, acceptable curves and unacceptable datasets respectively. Model with augmentation performs remarkably better than no augmentation. This is because enlarged datasets reduce overfitting and can effectively extract the image features. Furthermore, as shown in the Fig. 6, with the help of data augmentation, the training speed has slightly improvement where accuracy stabilized faster than no augmentation. Hence, by increasing the volume of datasets, overfitting can be significantly prevented, training speed and classification accuracy can be efficiently improved.

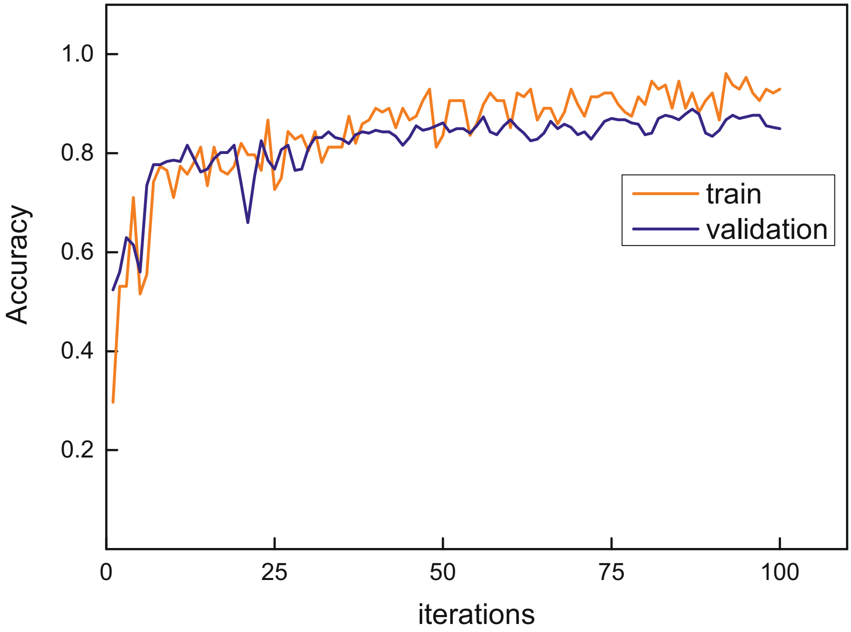


**Fig. 6.** Plot for accuracy on proposed model without data augmentation

### 3.2.3 Comparison with the State-of-the-Art Methods

To further verify the reliability and feasibility of our method, we compare the classification performance of our method with state-of-the-art methods. Through a massive number of literature and related technical research, inceptionV3 and Alexnet models have been chosen.

**Table 6.** Comparison of classification accuracy of different models

| Datasets | Our method | InceptionV3 | Alexnet |
|---|---|---|---|
| Precision | **0.9883** | 0.9732 | 0.8386 |
| AR | **0.9830** | 0.9713 | 0.7849 |
| TR | **0.9920** | 0.9716 | 0.9670 |
| UR | **0.9918** | 0.9771 | 0.7103 |

As can be seen from Table 6, it is evidently that our method generally achieves the best classification performance in overall datasets. The AR, TR, UR and precision measures tend to be significantly better than Alexnet method, and it is worth mentioning that the tendency of training and validation process of our method overwhelm that of Alexnet method. In addition, as we can see from Fig. 7, the accuracy curves of InceptionV3 and our method have overlapping parts during training and validation process. However,

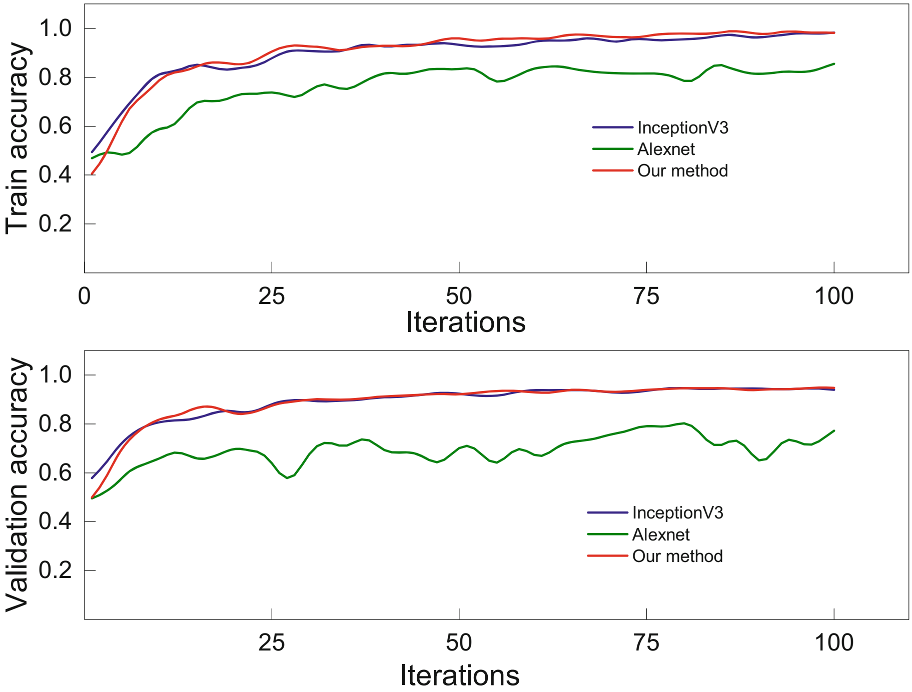


**Fig. 7.** Plot for accuracy on different methods

the results of our method show improvement in accuracy by 1.47% on training and validation.

### 3.3 Conclusion

In this paper, a convolution neural network combing with data augmentation is proposed for the classification of different types of make-up torque graphs in premium thread connections. The results show that using CNNs to predict and classify make-up torch graphs is proved to be a promising way to find the minor difference between typical curves, acceptable curves and unacceptable curves. Four specific measures namely, TR,AR, UR and Precision have been used to evaluate the performance of proposed method. The experimental results show that our model achieves the highest classification accuracies of 99%, 98%, 97%, 99% in presented measures respectively. Furthermore, with help of data augmentation techniques, the number of graphs in each dataset has been expanded which weaken the overfitting problem caused by lack of training data and improve the classification accuracy. Particularly, we would expect that such data augmentation techniques can be used to benefit the training work involved expert domain knowledge where collecting sufficient data is heavily hindered. Finally, comparison experiments have been conducted to verify the robustness of our method. The results confirms that proposed method outperforms on selected two state-of-the-art methods (InceptionV3, Alexnet).

In future studies, it is suggested to investigate more efficient strategies to improve the classification accuracy. Besides, the classified results of make-up torque curves are recommended to evaluate the sealing performance of premium thread connections. It is helpful to predict the sustained casing pressure condition.

## References

1. Guo, J., Ma, F.: Air tightness performance assessment of screw threads of oil tubings in high-sulfur gas wells in the Longgang Gas Field, Sichuan Basin. Nat. Gas Ind. **33**, 128–131 (2013)
2. Huang, Y., et al.: Safety assessment of production casing in HTHP CO_2 gas well. Drill. Prod. Technol. **37**, 78–81 (2014)
3. International N.: Laboratory Testing of Metals for Resistance to Sulfide Stress Cracking and Stress Corrosion Cracking in H2S Environments (2015)
4. Jiang, F., IN Lab: Evaluation methods of metal materials for high sour gas fields. Natural Gas Industry (2004)
5. Jianjun, W.: Seal test of the injection-production string for underground gas storage. China Pet. Machinery **42**, 170–173 (2014)
6. Jie, L., et al.: Technology and practice of well completion and putting into production for gas wells with high sulfur content in Sichuan and Chongqing areas. Nat. Gas Industry **26**(1), 72–75 (2006)
7. Li, Y., et al.: Completion difficulties of HTHP and high-flowrate sour gas wells in the Longwangmiao Fm gas reservoir, Sichuan Basin, and corresponding countermeasures. Nat. Gas. Ind. **3**(3), 269–273 (2016)
8. Liu, Y., et al.: An optimal design of pipe strings for horizontal sour gas wells at the Dawan Block, Puguang Gas Field. Nat. Gas Ind. **32**(12), 71–74 (2012)
9. Yinda, H.E., et al.: Analysis of tubing hermetic sealing in high pressure gas well of Tarim oilfield. Drill. Prod. Technol. **33**, 36–39 (2010)
10. Zhang, Z., et al., Wellbore integrity design of high-temperature gas wells containing CO_2. Natural Gas Industry, 2013
11. Chen, W., et al.: The sealing mechanism of tubing and casing premium threaded connections under complex loads. J. Petrol. Sci. Eng. **171**, 724–730 (2018)
12. Braswell, Gentry: Artificial intelligence comes of age in oil and gas. J. Pet. Technol. **65**(01), 50–57 (2013)
13. Jose, F., Carlos, R., Pedro, R.: An intelligent identification method of fuzzy models and its applications to inversion of NMR logging data. In: SPE Technical Conference & Exhibition (2002)
14. Al-Fattah, S.M., Startzman, R.A.: Predicting natural gas production using artificial neural network. In: SPE Hydrocarbon Economics & Evaluation Symposium
15. Mohaghegh, S.D.: Recent developments in application of artificial intelligence in petroleum engineering. J. Pet. Technol. **57**(04), 86–91 (2005)
16. Vikas, B., Hopkinson, S.: A novel approach to identify reservoir analogues. In: European Petroleum Conference (2002)
17. Bing, T., et al.: Video object detection for tractability with deep learning method. In: Fifth International Conference on Advanced Cloud & Big Data (2017)
18. Qin, J., et al.: A biological image classification method based on improved CNN. Ecol. Inform. **58**, 101093 (2020)
19. Sangineto, E., et al.: Self paced deep learning for weakly supervised object detection. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 712–725 (2016)
20. Yuan, C., et al.: Deep residual network with adaptive learning framework for fingerprint liveness detection. IEEE Trans. Cognit. Dev. Syst. **12**(99), 461–473 (2019)

21. Yuan, Y., Mou, L., Lu, X.: Scene recognition by manifold regularized deep learning architecture. IEEE Trans. Neural Netw. Learn. Syst. **26**(10), 2222 (2017)
22. Chan, T.H., et al.: PCANet: a simple deep learning baseline for image classification? IEEE Trans. Image Process. **24**(12), 5017–5032 (2015)
23. Licheng, et al.: Wishart deep stacking network for fast POLSAR image classification. IEEE Trans. Image Process. A Publication of the IEEE Signal Processing Society **25**(7), 3273–3286 (2016)
24. Zhang, J., et al., Lightweight deep network for traffic sign classification. Ann. Telecommun. Annales des télécommunications **75**(3), 369–379 (2019)
25. Ma, W., et al.: Adaptive median filtering algorithm based on divide and conquer and its application in CAPTCHA recognition. Comput. Mater. Continua **58**(3), 665–677 (2019)
26. Aka, B., et al.: Improved deep convolution neural network (CNN) for the identification of defects in the centrifugal pump using acoustic images. Appl. Acoust. **167,** 107399 (2020)
27. Shawky, O.A., et al.: Remote sensing image scene classification using CNN-MLP with data augmentation. Optik Int. J. Light Electron Opt. **221**, 165356 (2020)
28. Peng, Y.: Artificial intelligence applied in sucker rod pumping wells: intelligent dynamometer card generation, diagnosis, and failure detection using deep neural networks. In: SPE Annual Technical Conference and Exhibition (2019)
29. Sharaf, S.A., et al.: Beam Pump dynamometer card classification using machine learning. In: SPE Middle East Oil and Gas Show and Conference (2019)
30. Perez, L., Wang, J.: The Effectiveness of Data Augmentation in Image Classification using Deep Learning (2017)
31. LeCun, Y.A., Bottou, L., Orr, G.B., Müller, K.-R.: Efficient BackProp. In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) Neural Networks: Tricks of the Trade. LNCS, vol. 7700, pp. 9–48. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35289-8_3
32. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines Vinod Nair. In: International Conference on International Conference on Machine Learning (2010)
33. Hinton, G.E., et al.: Improving neural networks by preventing co-adaptation of feature detectors. Comput. Sci. **3**(4), 212–223 (2012)
34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Computer Science (2014)
35. Zeiler, M., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp.818–833. Springer, Cham (2013). https://doi.org/10.1007/978-3-319-10590-1_53
36. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vision **115**(3), 211–252 (2015)
37. Khandezamin, Z., Naderan, M., Rashti, M.J.: Detection and classification of breast cancer using Logistic Regression feature selection and GMDH classifier. J. Biomed. Inform. **111**, 103591 (2020)

# Construct a Drilling Complexity Intelligent Prediction Model Based on the Case-Based Reasoning

Hui-ying Zhai[1](✉), Bao-lin Liu[1], Ya-qiang Chen[2], and Cai-xia Lv[3]

[1] School of Engineering and Technology, China University of Geosciences, Wuhan, China
zhaihy@cugb.edu.cn
[2] China National Oil and Gas Exploration and Development Company, Beijing, China
[3] College of Robotics, Beijing Union University, Beijing, China

**Abstract.** In order to diagnose and predict the drilling complexities before drilling operations, and provide the drilling operators on well site with some hints, so that they are mentally aware of what kind of drilling complexity will occur in the future, and they could take the corresponding preventive measures to prevent the occurrence of some drilling complexities with the lowest possible economic cost. The handling methods of the drilling complexity cases that have occurred are treatment solutions made by the experts on well site based on their professional knowledge and years of rich drilling experience, which has a very important reference value for the later drilling operations. On the basis of case-based reasoning method, according to the adjacent well data, the computer technology, the artificial intelligence, and the data mining technology, this paper will construct a drilling complexity intelligent prediction model to utilize an open-source software and regression analysis method of causal relationship model. Use the ROC curve and confusion matrix to evaluate the performance of the drilling complexity intelligent prediction model, and the accuracy of the model is between 70–80%. It is recommended to use more drilling complexity cases to train the model in the later stage to improve the accuracy of the prediction model.

**Keywords:** Drilling Complexity · Case-based Reasoning · Python · Intelligent Prediction Model · ROC curve · Confusion Matrix

## 1 Introduction

Through extensive literature research, it can be seen that drilling complex warning systems generally use real-time drilling parameters for warning, but the transmission of drilling parameters has a lag. Therefore, the complex situations predicted by warning systems that rely on drilling parameters have actually occurred underground. Moreover, due to the complexity and diversity of the formation, it is difficult to predict the occurrence of all drilling complex situations.

This paper attempts to use the data of drilled wells, computer technology, artificial intelligence and data mining technology to diagnose and predict the drilling complexity

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 346–352, 2024.
https://doi.org/10.1007/978-981-97-0272-5_30

before well operation under the condition of reducing economic costs, so as to give hints to the operators on well site, so that they could realize in advance what kind of drilling complexity will occur, take corresponding preventive measures, prepare sufficient materials, tools and equipment, find reasonable solutions to prevent drilling complex situations from occurring at the lowest possible time and economic cost, or handle drilling complex timely.

The modern oil drilling operation process is generally recorded, and the information for each well includes well design, drilling daily report, mud daily report, drilling cost daily report, casing data, well deviation data, completion report, logging data, etc.

## 2 The Development of Intelligent Diagnosis and Prediction Technology

With the rapid development of computer, artificial intelligence, data mining and other technologies, prediction technology has also undergone rapid development, and various intelligent computing methods have emerged and been applied to the field of prediction. Artificial intelligence technology will play an increasingly important role in the field of prediction. Artificial intelligence is the use of more complex technologies to replace human brain decision-making, by searching and analyzing in databases, and building models. These are not repetitive tasks, but require judgments based on complex algorithms and machine learning, which can be applied to predict future development trends and make scientific and reasonable decisions. When artificial intelligence is applied to prediction, the algorithm is the key, the data and computing power are the foundation. The widespread application of this technology is due to the accumulation of data and computational power. At present, in terms of economy (stock prices, various competition results), healthcare, agriculture, etc., models are established and machine learning algorithms are used to predict future results. At present, the methods applied to intelligent diagnosis and prediction mainly include Artificial neural network, support vector machine, Chaos theory, Swarm intelligence (i.e. particle swarm optimization), Wavelet analysis, Time series, Case-based reasoning, Expert system, etc.

## 3 Case-Based Reasoning

Case-based reasoning (CBR) is both an artificial intelligence technique and an analogical reasoning method that utilizes relevant information from similar cases to solve current problems.

CBR is the process of finding similar historical cases and utilizing specific knowledge from existing experience or results, i.e. specific cases, to solve new problems.

CBR technology originated from Yale University in the United States, and was described by Roger Schank in Dynamic Memory in 1982. It is an important knowledge-based problem-solving and learning method that has emerged in the field of artificial intelligence. It solves problems by reusing or modifying previous solutions to similar problems. The CBR research method originates from human cognitive and psychological activities, alleviating the bottleneck problem of knowledge acquisition in conventional knowledge systems. It combines quantitative analysis with qualitative analysis, and has the characteristics of dynamic knowledge base and incremental learning.

The basic steps of a typical CBR problem-solving process can be summarized into four main processes: case retrieval, reuse, revision, and retention. Therefore, CBR is also known as 4R. In CBR, the problem or condition to be solved is usually referred to as the target case, historical cases are referred to as the base case, and the collection of source cases is referred to as the case library. From this, it can be understood that the basic process of CBR problem-solving is: a new problem that needs to be solved arises, and this is the target case; Using the description information of the tar-get case to query similar past cases, that is, searching the case library to obtain source cases that are similar to the target case, and thus obtaining some solutions to new problems; If this solution fails, adjustments will be made to obtain a successful case that can be saved. After this process is completed, a more complete solution for the target case can be obtained; If the source case fails to provide a correct and appropriate solution, a new source case can be obtained by correcting and saving the case. In the process of CBR, case representation, case retrieval, and case adjustment are the core issues of CBR research.

# 4 Constructing a Drilling Complex Intelligent Prediction Model

This paper uses Python language for data analysis and construction the prediction model.

## 4.1 Process of Model Construction

The specific process of model construction is shown in Fig. 1.

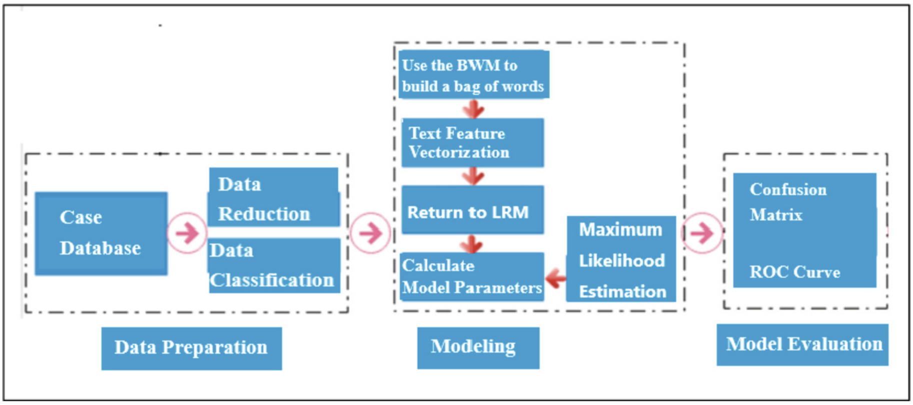


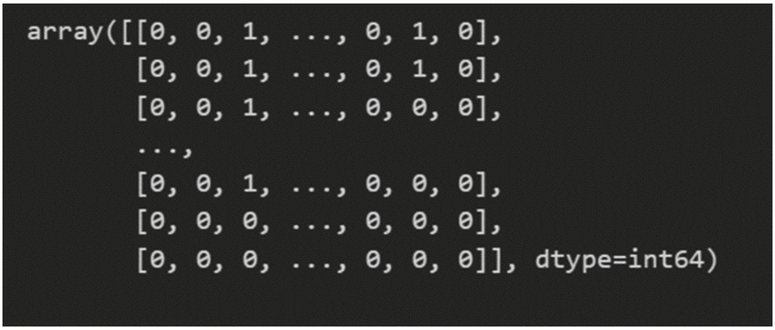
**Fig. 1.** Flow Chart for Constructing Drilling Complex Intelligent Prediction Model. BWM—Bag-of-words model; LRM—Logistic regression model.

## 4.2 Model Construction

Based on the drilling data of the oilfield, extract and organize complex cases of drilling encountered in the oilfield, and form a case library. Using the data from the drilling complex case library, based on statistical analysis, classify and simplify the data, while cleaning the data to ensure the effectiveness of the mapping relationship be-tween the data.

Using quantitative prediction method for quantitative analysis, that is, predicting the probability of various drilling complexities based on data such as drilling formations and lithology, and then predicting complex types based on specific geological conditions.

Logistic regression, machine learning algorithm and Python development language are used to build an intelligent diagnosis prediction model (Fig. 2). First, we use the Bag-of-words model model to vectorize the text features such as stratum and reason, and transform the Chinese text into a sentence vector that can be recognized by the computer (Fig. 3). Replace the complexity of label drilling with a scalar, such as severe leakage of 0, severe shrinkage of 1, sticking of 2, collapse of 3, overflow of 4, etc. The trans-formed sentence vector is transferred into the Logistic regression model, and the parameters of the Logistic regression model are calculated by maximum likelihood estimation. Finally, the probability of each drilling situation is obtained.



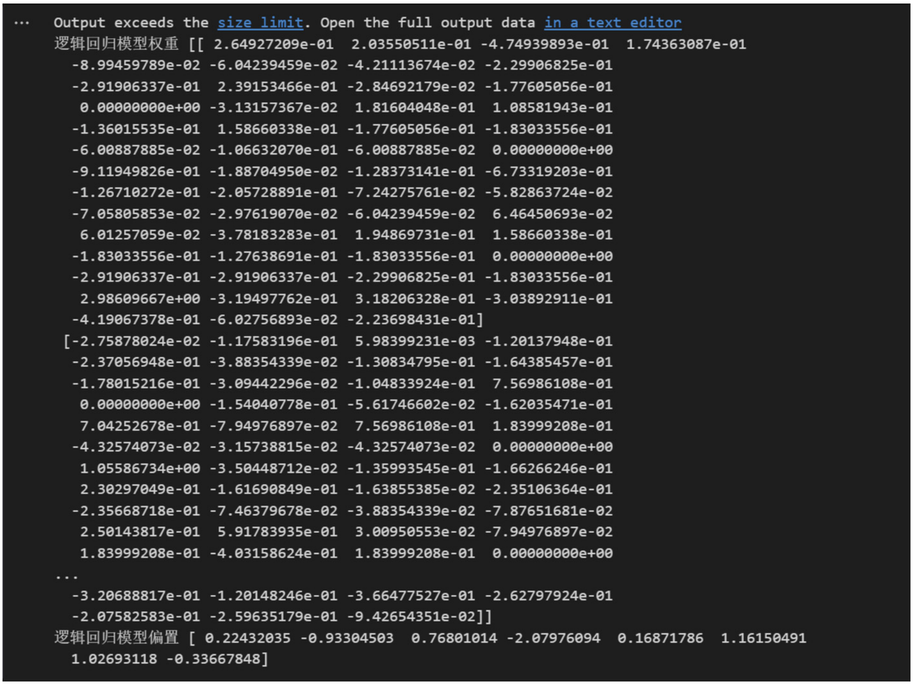
```
array([[0, 0, 1, ..., 0, 1, 0],
       [0, 0, 1, ..., 0, 1, 0],
       [0, 0, 1, ..., 0, 0, 0],
       ...,
       [0, 0, 1, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

**Fig. 2.** Intelligent diagnosis and prediction model

## 4.3 Model Evaluation

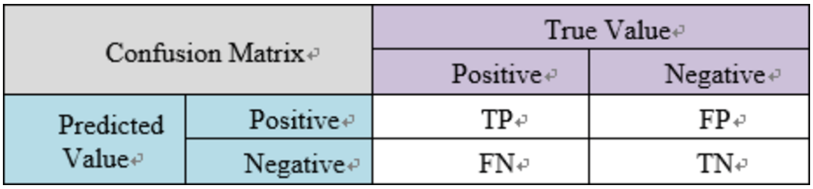
The Confusion matrix and Receiver operating characteristic are used to evaluate the performance of the model.

The Confusion matrix is a situation analysis table that summarizes the prediction results of the classification model in machine learning. The records in the dataset are summarized in the form of a matrix according to the two criteria of the real category and the category judgment predicted by the classification model, as shown in Fig. 4. The columns of the matrix represent the true values, and the rows of the matrix represent the predicted values, as shown in Fig. 5. In the Confusion matrix, the higher the value at the corresponding position of TP and TN, and the lower the value at the corresponding position of FP and FN, the higher the accuracy of the model, which can be calculated by formula (1). After calculation, the accuracy of the model in predicting drilling complexity is around 78% (Fig. 6).



```
···  Output exceeds the size limit. Open the full output data in a text editor
     逻辑回归模型权重 [[ 2.64927209e-01  2.03550511e-01 -4.74939893e-01  1.74363087e-01
       -8.99459789e-02 -6.04239459e-02 -4.21113674e-02 -2.29906825e-01
       -2.91906337e-01  2.39153466e-01 -2.84692179e-02 -1.77605056e-01
        0.00000000e+00 -3.13157367e-02  1.81604048e-01  1.08581943e-01
       -1.36015535e-01  1.58660338e-01 -1.77605056e-01 -1.83033556e-01
       -6.00887885e-02 -1.06632070e-01 -6.00887885e-02  0.00000000e+00
       -9.11949826e-01 -1.88704950e-02 -1.28373141e-01 -6.73319203e-01
       -1.26710272e-01 -2.05728891e-01 -7.24275761e-02 -5.82863724e-02
       -7.05805853e-02 -2.97619070e-02 -6.04239459e-02  6.46450693e-02
        6.01257059e-02 -3.78183283e-01  1.94869731e-01  1.58660338e-01
       -1.83033556e-01 -1.27638691e-01 -1.83033556e-01  0.00000000e+00
       -2.91906337e-01 -2.91906337e-01 -2.29906825e-01 -1.83033556e-01
        2.98609667e+00 -3.19497762e-01  3.18206328e-01 -3.03892911e-01
       -4.19067378e-01 -6.02756893e-02 -2.23698431e-01]
      [-2.75878024e-02 -1.17583196e-01  5.98399231e-03 -1.20137948e-01
       -2.37056948e-01 -3.88354339e-02 -1.30834795e-01 -1.64385457e-01
       -1.78015216e-01 -3.09442296e-02 -1.04833924e-01  7.56986108e-01
        0.00000000e+00 -1.54040778e-01 -5.61746602e-02 -1.62035471e-01
        7.04252678e-01 -7.94976897e-02  7.56986108e-01  1.83999208e-01
       -4.32574073e-02 -3.15738815e-02 -4.32574073e-02  0.00000000e+00
        1.05586734e+00 -3.50448712e-02 -1.35993545e-01 -1.66266246e-01
        2.30297049e-01 -1.61690849e-01 -1.63855385e-02 -2.35106364e-01
       -2.35668718e-01 -7.46379678e-02 -3.88354339e-02 -7.87651681e-02
        2.50143817e-01  5.91783935e-01  3.00950553e-02 -7.94976897e-02
        1.83999208e-01 -4.03158624e-01  1.83999208e-01  0.00000000e+00
     ...
       -3.20688817e-01 -1.20148246e-01 -3.66477527e-01 -2.62797924e-01
       -2.07582583e-01 -2.59635179e-01 -9.42654351e-02]]
     逻辑回归模型偏置 [ 0.22432035 -0.93304503  0.76801014 -2.07976094  0.16871786  1.16150491
       1.02693118 -0.33667848]
```

**Fig. 3.** Text feature vectorization model

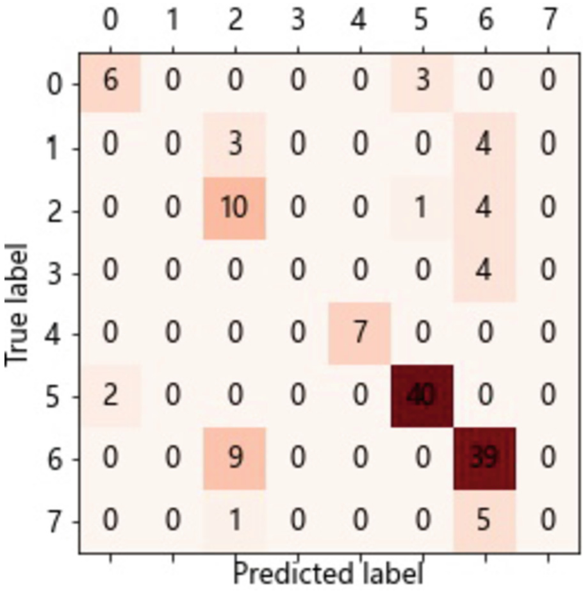


| Confusion Matrix | | True Value | |
|---|---|---|---|
| | | Positive | Negative |
| Predicted Value | Positive | TP | FP |
| | Negative | FN | TN |

**Fig. 4.** Sketch map of confusion matrix

$$(\text{Accuracy}) = (TP + TN) / (TP + FP + FN + TN) \tag{1}$$

ROC (Receiver Operating Characteristic) curve was invented by radar engineers in World War II to detect enemy vehicles (aircraft and ships) on the battlefield, that is, Detection theory. Later, it was introduced into the field of machine learning to judge the quality of classification and detection results. AUC (Area Under Curve) refers to the area under the Receiver operating characteristic, which is used to measure the classification effect of the model. It is an evaluation indicator to measure the quality of the model, and represents the probability that positive cases rank ahead of negative cases. After calculation, AUC under the Receiver operating characteristic of the intelligent diagnosis and prediction model for the drilling is 68%.



| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| 0 | 6 | 0 | 0 | 0 | 0 | 3 | 0 | 0 |
| 1 | 0 | 0 | 3 | 0 | 0 | 0 | 4 | 0 |
| 2 | 0 | 0 | 10 | 0 | 0 | 1 | 4 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 |
| 4 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 |
| 5 | 2 | 0 | 0 | 0 | 0 | 40 | 0 | 0 |
| 6 | 0 | 0 | 9 | 0 | 0 | 0 | 39 | 0 |
| 7 | 0 | 0 | 1 | 0 | 0 | 0 | 5 | 0 |

**Fig. 5.** Confusion matrix vector diagram

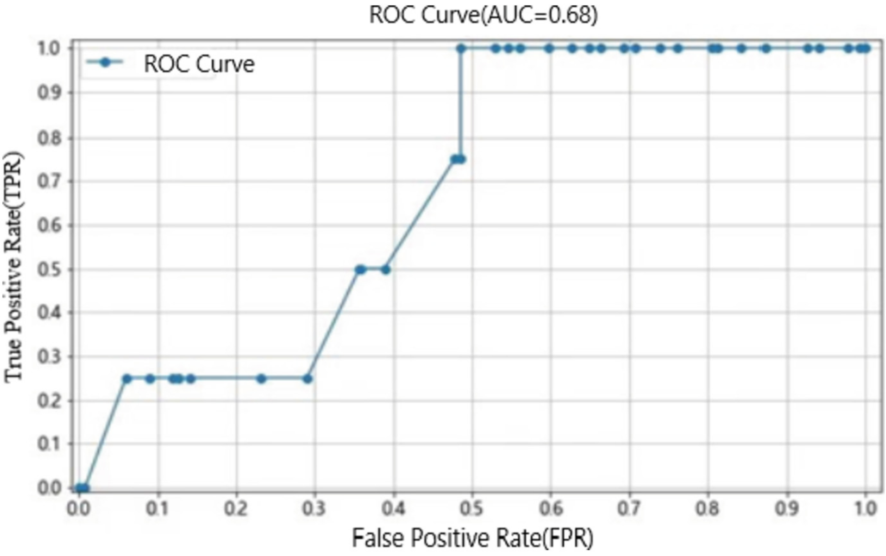


**Fig. 6.** ROC evaluation curve of prediction mode

## 5 Conclusion and Suggestion

1) Building a prediction model generally consists of three steps, namely data preparation, selecting appropriate methods to build the prediction model, and model evaluation.

2) Based on the drilling complex case database, using Python language, Logistic regression analysis, maximum likelihood estimation and machine algorithm, a drilling complex intelligent diagnosis and prediction model for one oil field is constructed.
3) With the increase of data from the oilfield, it is recommended to use more com-plex drilling cases to train the model to improve the accuracy of model prediction.

## References

1. Roger, C.S.: Dynamic Memory: A Theory of Reminding and Learning in Computers and People. Cambridge University Press, Cambridge (1982)
2. Yost, K., Valentin, A., Einstein, H.H.: Estimating cost and time of wellbore drilling for Engineered Geothermal Systems (EGS) – Considering uncertainties. Geothermics **53**, 85–99 (2015)

# Soft Actor-Critic Based Deep Reinforcement Learning Method for Production Optimization

Guo-jing Xin[1], Kai Zhang[1,2](✉), Zhong-zheng Wang[1], Zi-feng Sun[1], Li-ming Zhang[1], Pi-yang Liu[2], Yong-fei Yang[1], Hai Sun[1], and Jun Yao[1]

[1] College of Petroleum Engineering, China University of Petroleum (East China), Qingdao, China
zhangkai@upc.edu.cn

[2] School of Civil Engineering, Qingdao University of Technology, Qingdao, China

**Abstract.** Production optimization is a crucial technology for efficient development of water-driven reservoirs. By adjusting the injection and production rate of oil and water wells in a reservoir block, the optimal production solution can be provided for the field to maximize the economic benefits while minimizing the costs. In this paper, a soft actor-critic (SAC) based reinforcement learning offline production optimization method are proposed, which models the production optimization problem as a Markov sequence decision process. Specifically, the deep reinforcement learning (DRL) agents aimed at maximizing the economic efficiency. The agent updated the policy model incrementally using the data obtained by interactive sampling with the environment to accelerate the convergence of the optimization process. In addition, to achieve offline optimization, a state transfer model is constructed that captures the dynamics of the reservoir under time-varying well control conditions using historical regulation experience. In the offline deployment stage of the cloud platform, the trained policy network and state transition network are utilized. In this way, the well control scheme for multiple future time steps can be calculated using only the current state of the reservoir. Reservoir instances show that this method is highly efficient and can provide optimized solutions within seconds, and the optimization performance is also remarkable. With the good effect of water control and oil increment, the target model can achieve higher net present value (NPV). The proposed offline method,

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 353–366, 2024.
https://doi.org/10.1007/978-981-97-0272-5_31

which embedding control strategies into the model and utilizing a state transition model to capture the dynamics of the system, offers a novel approach to intelligent production optimization. By enabling offline optimization deployment on a cloud platform, this approach provides a practical solution to meet the demand for intelligent oilfield construction.

**Keywords:** Water Driven Reservoirs · Production Optimization · Deep Reinforcement Learning · Policy Model · Intelligent Oilfield

## 1 Introduction

The new round of oil and gas technology revolution and digital revolution is sweeping the world with unprecedented breadth and depth, and the cross-border integration of new technologies such as the Internet of Things, artificial intelligence, and 5G with the oil and gas industry has become an important way of innovation. The construction of intelligent oil and gas fields has become an inevitable trend, with intelligent production optimization being a key link.

Production optimization of oil reservoir is a common means to achieve increased and stable production in oil fields. By reasonably optimizing the injection and production schedule of oil and water wells, the development effect is improved, achieving the effect of minimizing water production while increasing oil recovery. After recent development, mainstream production optimization methods mainly include three categories: gradient based algorithms, stochastic optimization algorithms, and machine learning methods.

Gradient based methods were first proposed in various research works. In 2004, Brouwer et al. [1] combined integrated Kalman filter and concomitant algorithm to improve NPV for water-driven reservoir development. Zhang et al. [2] proposed a combination of finite difference and stochastic gradient method to solve constrained reservoir production optimization problems in 2016.Unlike gradient-based methods, stochastic optimization methods do not need to compute gradients explicitly and have better flexibility. In 2018, Foroud et al. [3] applied several global optimization algorithms to the Brugge field and compared their optimization performance. In 2022, Yin et al. [4] proposed a multi-fidelity optimization framework, which differential evolution algorithms and numerical simulators with different fidelity are combined. Machine learning surrogate model is a combination of machine learning model and optimization algorithm, and its core idea is to construct a machine learning proxy model instead of reservoir numerical simulation in the real-time optimization process, which can significantly reduce the time cost of iterative computation. In 2021, Zhang et al. [5] used a support vector machine as an surrogate model. Chen et al. [6] developed a radial basis function-based surrogate-assisted evolutionary algorithm to solve a high-dimensional expensive production optimization problem in 2022.

However, the above methods face the inability to truly embed the regulation experience in reservoir production and development. It relies only on the fitness value for the adjustment of the system. Additionally, each application requires multiple interactions with the numerical simulator, making it difficult to achieve offline cloud applications.

In this paper, we investigate the above problems and introduce reinforcement learning theories and methods, which have been successfully applied in several fields. In 2019, Ma et al. [7] applied various deep reinforcement learning methods to optimize the injection rate of water-driven reservoirs. The results showed that deep reinforcement learning methods converge faster than traditional methods in some cases. In Wang et al. [8–10] proposed the use of reinforcement learning method for the life-cycle production optimization of oil reservoirs.

This paper proposes an offline production optimization method that utilizes the empirical data accumulated during the trial and error process of reinforcement learning. During the policy training process, the empirical data is utilized to train an additional state transition model to achieve multi-step offline applications on cloud platforms. During the optimization stage, the offline method can break away from numerical simulation and provide injection and production of oil and water wells for future multiple time steps only based on the current state.

This paper proceeds as follows. In Sect. 2, we present production optimization for subsurface oil-water flow and briefly describe theories related to machine learning. In Sect. 3, the offline method is presented, the agent and state transition models are highlighted. The page presents the results for two-dimensional oil-water problems in Sect. 4. Test cases involve the three channels model and a gaussian model. Then, the summary and suggestions for future work are concluded in Sect. 5.

## 2 Methodology

### 2.1 Development of Mathematical Model for the Production Optimization

With the development of oil fields, most of the oil fields in China are now in the period of high water content, so it is necessary to regulate the production system to improve the field revenue. The production optimization problem is to achieve the goal of maximizing economic efficiency or oil production by changing the production rate and injection rate in an oilfield block. In this process, the reservoir is usually considered as a complex dynamic system, and a mathematical model for reservoir development and production optimization is established with economic net present value (NPV) as the objective. The formula for calculating the NPV is shown as

$$NPV_t = c_1 P_t^{\mathrm{oil}} - c_2 P_t^{\mathrm{water}} - c_3 I_t^{\mathrm{water}} \tag{1}$$

where $P_t^{oil}$ is the cumulative oil production during the time period, STB; $P_t^{water}$ is the cumulative water production during the time period, STB; $I_t^{water}$ is the cumulative oil injection during the time period, STB; $c1$ is the oil revenue, USD/STB; $c2$ is the cost of disposing produced water, USD/STB; $c3$ is the cost of injecting water, USD/STB.

### 2.2 Deep Reinforcement Learning

Deep reinforcement learning constitutes a significant research domain in the field of artificial intelligence, with potential enormous application value. The remarkable achievements of intelligent agents represented by Alpha Zero also indirectly demonstrates their ability to solve complex high-dimensional problems. Reinforcement learning finds widespread use across diverse fields, including but not limited to gaming, autonomous driving, and robot route planning.

Reinforcement learning is a solution to the problem of intelligent agents maximize cumulative rewards in complex and uncertain environments, which mainly includes two parts: the agent and the environment, as shown in Fig. 1. Intelligent agents can achieve direct control from raw input to output through end-to-end learning.

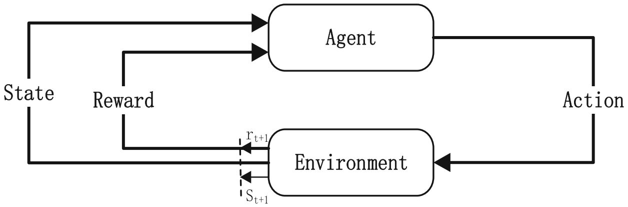


**Fig. 1.** Schematic diagram of reinforcement learning flow

In this paper, a soft-actor-critic framework is used to build the model. Soft Actor Critic (SAC) algorithm is an off-policy algorithm proposed by Haarnoja [11, 12] in 2018. SAC is a reinforcement learning algorithm that blends stochastic policy optimization with deep deterministic policy gradient techniques. A key characteristic of SAC is the use of entropy regularization, which maximizes the trade-off between expected return and policy entropy.

### 2.3 Deep Learning

Convolutional neural network (CNN) learns from the information processing process of the human brain that extracted features from the data layer by layer. The features extracted through the progressive number of layers will become more complex. Feature extraction of image data is achieved through local connections and weight sharing. Similarly, deep convolutional neural networks have been widely applied in the petroleum field, such as residual oil distribution prediction [13, 14], history matching [15, 16], production optimization, etc. In this paper, convolutional neural networks are used to extract state information features.

## 3 Off-Line Model Design for Production Optimization

### 3.1 Environment Model

In the production optimization problem, the environment model is built with the help of the reservoir numerical simulator ECLISPE. After the agent generates the action, the reservoir numerical simulator simulates according to this well control schedules to reflect the actual production state of the reservoir. In this paper, the black oil model is used for the flow simulation of oil and water phases. Governing partial differential equations and constraints are (2)–(5).

$$-\frac{k_{r,\alpha}k}{\mu_\alpha}(\nabla P_\alpha - \rho_a g \nabla z) = q_\alpha \tag{2}$$

$$\nabla \cdot (q_\alpha) + q_{v,\alpha} = \frac{\partial(\varphi\rho_\alpha S_\alpha)}{\partial t} \tag{3}$$

$$P_c = P_o - P_w \tag{4}$$

$$S_o + S_w = 1 \tag{5}$$

where $\alpha = \mathrm{o}$ or w denote the oil or water phase; k is the absolute permeability and $k_{\mathrm{r},\alpha}$ is the relative permeability to phase $\alpha$; $P_\alpha$ is the fluid pressure for phase $\alpha$; $P_c$ denotes the capillary pressure; $q_\alpha$ is the phase flux; $S_\alpha$ are phase saturation; z is the reservoir depth; $g$ is the acceleration of gravity; $\Phi$ is the rock porosity.

## 3.2 Agent Model

When applying reinforcement learning to reservoir development and production scenarios, there are several points to note. The parameters of the reservoir, such as saturation, pressure, is called the "state". It is part of the environment that the agent is observing and serves as input. The variables that the problem needs to optimize are called "action", such as reservoir parameters or production systems. After adjusting the parameters, the state of the model changes, it reaches the "next state". At the same time, the environment gives feedback to the corresponding actions, which are converted into rewards for the agent, known as "rewards." By continuously interacting with the environment, the reinforcement learning method learns how to make decisions that maximize the reward, called "policy". The policy is the core of the offline model. Figure 2(a) shows the network structure of agent, including actor and critic parts.

**Action Space**

In the production optimization problem, the optimization variable is the production system of each well in the reservoir. Therefore, in the design of the RL model for production optimization, the action space of the agent is shown as the following (6), and the action space is represented as an N-dimensional vector.

$$a_t = \left\{BHP_{prd,1}, \cdots, BHP_{prd,N_{prd}}; w_{inj,1}, \cdots, w_{inj,N_{nj}}\right\} \tag{6}$$

where, $BHP_{prod,i}$ represents the bottom hole pressure of producing well, $w_{inj,i}$ represents the injection rate from the injection well.

During the actual development process, the range of each well should be designed by field engineer. The range, which can be called the linear constraint, is transformed using a linear mapping of the network output values. That is, the terms in the Eq. (6) need to satisfy the (7) conditions.

$$\begin{aligned} x \le BHP_{prd,i} \le y \\ m \le w_{inj,j} \le n \end{aligned} \tag{7}$$

where x and y are upper and lower limits of bottom hole flow pressure, m and n is upper and lower limits of injection rate. $BHP_{prod,i}$ is the bottom hole pressure of well $i^{th}$, $w_{inj,j}$ is the injection rate of well $j^{th}$.

**State Space**

The state space of the agent includes the reservoir saturation and pressure. The state space can be expressed as shown in Eq. (8). Where the field information is input in the form of a matrix whose size is consistent with the reservoir grid size.

$$S_t = \left\{ s_{\mathrm{w,t}}, p_t \right\} \tag{8}$$

where $S_t$ means the state at time t, $s_{w,t}$ is the water saturation at time t, $p_t$ is the reservoir pressure at time t.

**Reward Function**

Reward function is a scalar feedback signal given by the environment that is intended to respond to the effect of an action taken by the agent in a given state. According to the previous section, the purpose of DRL is to maximize the cumulative reward. In the traditional production optimization problem, the objective of the study is to maximize the net present value NPV, therefore, the reward function is set to the Eq. (1) in this paper.

### 3.3 State Transition Model

The output of the agent depends on the observed state information, but the actual cloud platform cannot carry out the relevant calculation of reservoir numerical simulation. Therefore, the convolutional neural network is adopted to establish the state transition prediction model, whose input is the current state St and the well control At, and the output is the state $S_{t+1}$ at the next time, so as to realize the state transition away from the numerical simulation. Figure 2(b) shows the network structure of the state transition model.

### 3.4 Network Structure and Algorithm Flow

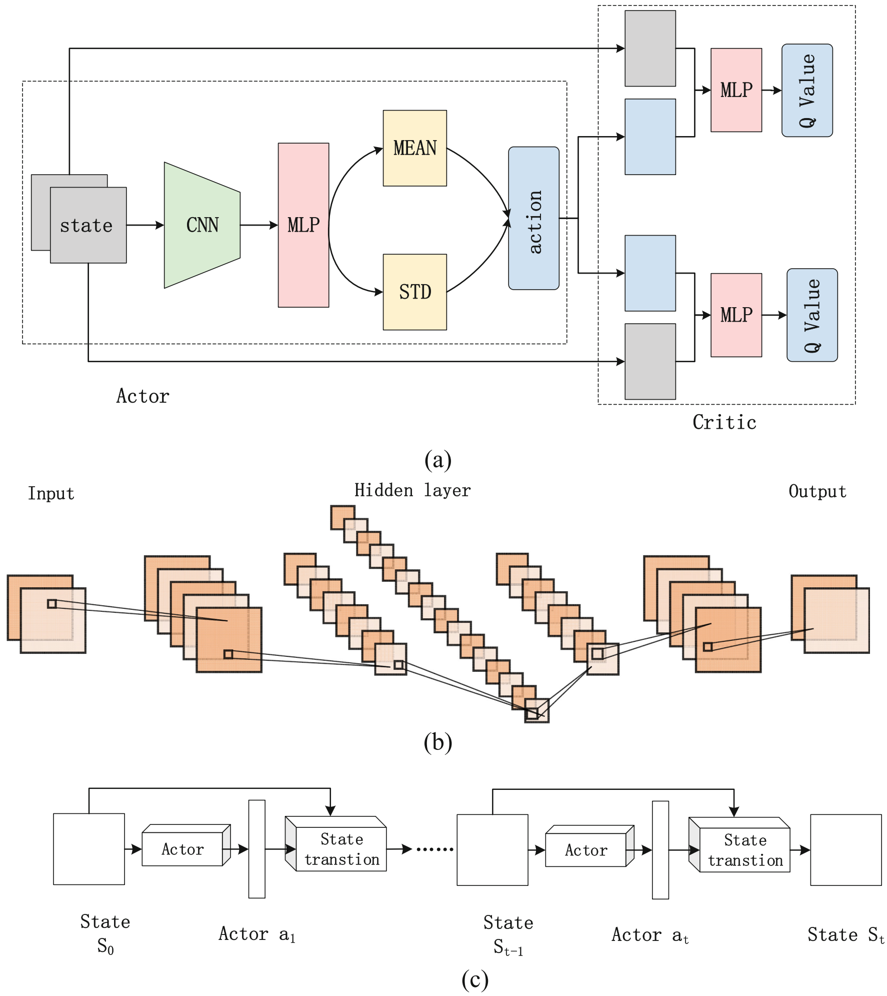
According to the above description of the agent model and state transition model, the algorithm can be divided into the training phase and the application phase.

In the training phase, the optimal network model is obtained through interaction with the numerical simulator. The training process consists of a series of steps that involve initializing the agent, allowing the agent to interact with the environment, updating the agent's policy and state transition model, and repeating these steps until convergence or a stopping criterion is reached. The training process is as Table 1.

**Table 1.** Algorithmic pseudocode.

```
Algorithm: Pseudocode of offline production optimization based on SAC
Initialize policy parameter $\phi$; action-value function parameters$\theta_1, \theta_2$; state transition model parameter $\varphi$; $\mathcal{D} \longleftarrow \emptyset$;
Set up the reservoir simulation environment;
for each iteration do
  for each environment timestep do
    if $\mathcal{D} \geq start\_num$
      $a_t \sim \pi_\phi(a_t|S_t)$
    else
      $a_t = Random(a_t)$
    $S_{t+1} \sim p(S_{t+1}|a_t, S_t)$
    $\mathcal{D} \longleftarrow \mathcal{D} \cup \{S_t, a_t, r_t, S_{t+1}\}$
  end for
  for each gradient step do
    Randomly sample a mini-batch of transitions $(S_t, a_t, r_t, S_{t+1})$ from $\mathcal{D}$
    Update Q network parameters $\theta_i \longleftarrow \theta_i - \lambda_Q \nabla_{\theta_i} J_Q(\theta_i) \quad for\ i \in \{1,2\}$
    Update policy network parameters $\phi \longleftarrow \phi - \lambda_\pi \nabla_\phi J_\pi(\phi)$
    Update target network parameters $\bar{\theta}_i \longleftarrow \tau\theta_i + (1-\tau)\bar{\theta}_i \quad for\ i \in \{1,2\}$
    Update temperature parameter $\alpha \longleftarrow \alpha - \lambda \nabla_\alpha J(\alpha)$
  end for
  for each training step do
    Update state transition model parameter $\varphi \longleftarrow \varphi - \lambda_P \nabla_\varphi J_P(\varphi)$
  end for
end for
```

Once the agent's policy has been trained, it can be evaluated on a separate set of environments to gauge its performance. In the offline application stage, only the current state of the reservoir can realize the injection and production well control design for multiple time steps in the future. Figure 2(c) shows the process of the method's application.



(a)

(b)

(c)

**Fig. 2.** The network structure described above. (a) is the agent structure, the output of actor is the action and the output of critic net is Q value(b) is the network structure of the state transition model, the net input is state at t step and the output is state at t + 1;(c) is the offline application process.

# 4 Case Study and Results

## 4.1 Case 1: Three-Channel Reservoir Model

The three-channel model is used to test the effect of the two-dimensional reservoir model. The reservoir includes 4 water injection wells and 9 production wells, which are arranged according to the five-point well pattern. The reservoir thickness is 20ft, the model grid distribution is 25 × 25 × 1, and the grid size is $\Delta x = \Delta y = 100$ ft. Figure 3 shows the permeability field and well location distribution of the model. Other basic properties of the model are listed in Table 1 (Table 2).

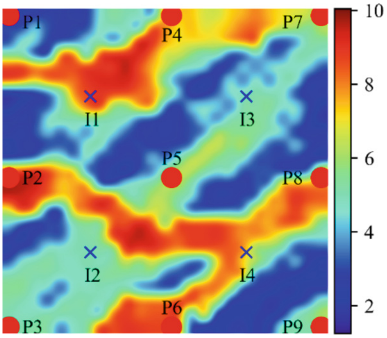


**Fig. 3.** Log permeability field and well location distribution of three-channel reservoir model.

**Table 2.** Parameters of three-channel reservoir model.

| Reservoir property | Parameter |
|---|---|
| Reservoir depth | 4800 ft |
| Porosity | 0.2 |
| Compression factor | $6.9 \times 10^{-5}$ psi$^{-1}$ |
| Crude oil viscosity | 2.2 cP |
| Initial water saturation | 0.2 |

Using the three-channel model, the classical GPEME algorithm is as a benchmark. The GPEME algorithm is an evolutionary optimization technique designed to address medium-scale, computationally expensive problems, which leverages a surrogate model based on the Gaussian process.

Figure 4 shows that the well control schedules calculated by using the state transfer model are similar to those calculated by directly interacting with the reservoir numerical simulator. Their calculated results reflect that the reservoir pressure is kept high for PRO-02 and PRO-05, and the injection volume of INJ-01 and INJ-02 is increased, thus proving that the application of the state transfer model can replace the reservoir numerical simulation to a certain extent.

**Table 3.** Calculation result

| | FOPT/STB | FWIT/STB | FWPT/STB | TIME |
|---|---|---|---|---|
| GPEME | 1404253 | 5780044 | 4951212 | – |
| Policy-ECL | 1514331 | 5242320 | 4399635 | 27.89 |
| Policy-Surrogate | 1508122 | 5331809 | 4508238 | 0.55 s |

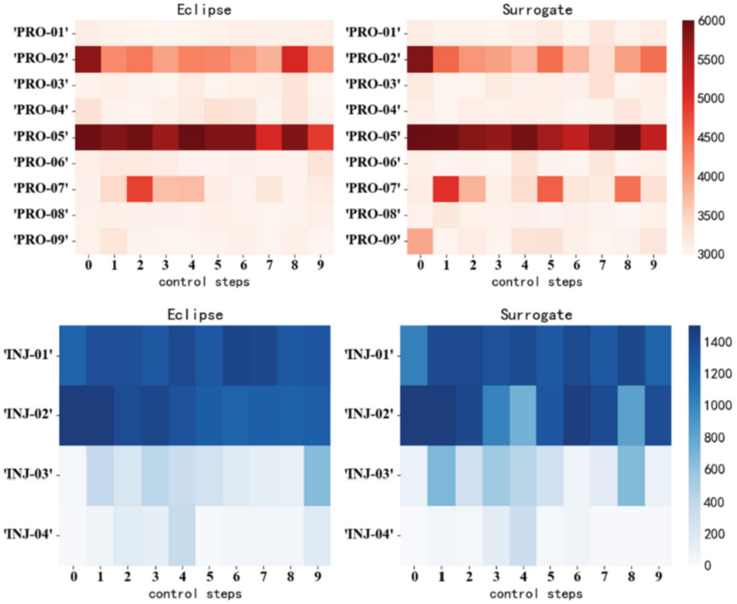


**Fig. 4.** Comparison of injection BHP and production rate (three-channel reservoir model). On the left are the results of interaction with the numerical simulator, and on the right are the results with the help of the state transition model, the bottom hole pressure of the production well is at the top, and the injection well injection rate is at the bottom.

This paper compares approaches that involve interactions with a numerical simulator to those that do not, then the final generated well control schedules of both methods are put into the numerical simulator for calculation. It can be seen that the cumulative oil production of both methods is similar, but the proposed method in this paper has a reduced calculation time when applied, from 27.89 s to 0.55 s. Meanwhile the method detached from the numerical simulation can realize the application of cloud platform.

The generated well control schedules were put into the numerical simulation model for calculation, and the resulting water saturation fields are shown in Fig. 5, From Fig. 5, it can be seen that the oil displacement effect achieved by two methods is similar.

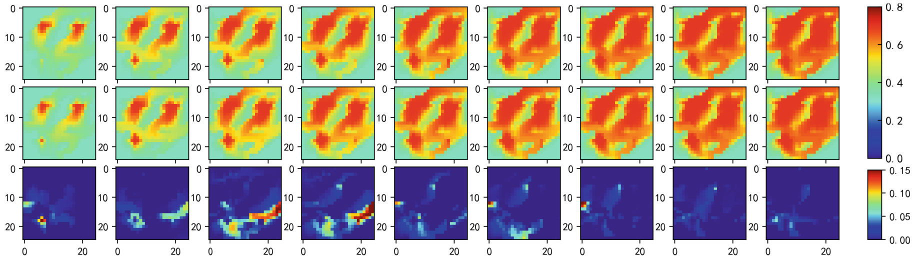


**Fig. 5.** Water saturation of three-channel reservoir model. The first row corresponds to the results of the injection and recovery regimes interacted with Eclipse, the second row corresponds to the results of the injection and recovery regimes corresponding to the state transfer agent model, and the third row corresponds to the difference between the two.

## 4.2 Case 2: Gaussian Reservoir Model

In Case2, gaussian model was adopted to test the effect. The reservoir size was 60*60, and the permeability field was generated by SGeMS. The reservoir consists of four injection Wells and nine production Wells arranged in a five-point pattern. Figure 6 shows the permeability field and well location distribution of the model. The basic attributes of the model are shown in Table 3. The optimization cycle is 1800 days, 180 days as a step, a total of 10 steps.

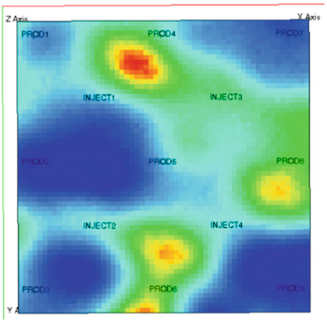


**Fig. 6.** Permeability field and well location distribution of gaussian reservoir model.

Table 4 and Fig. 7, Fig. 8 are the result of gaussian model. It shows that the calculation time reduced from 64.16s to 3.38 s. Because there is no need to perform complex numerical simulation calculations. Table 4 also shows that FOPT (field oil production total) is slightly higher, but FWIT (field water injection total) and FWPT (field water production total) are also slightly higher. That indicates that the increase of the injection rate leads to the increase of the production.

**Table 4.** Parameters of gaussian model.

| Reservoir property | Parameter |
|---|---|
| Mesh size | 60*60 |
| Reservoir depth | 4005 m |
| Porosity | 0.2 |
| Crude oil viscosity | 2 cp |

**Table 5.** Calculation result.

| | FOPT/STB | FWIT/STB | FWPT/STB | TIME |
|---|---|---|---|---|
| GPEME | 1467477 | 2175852 | 612316.2 | – |
| Policy-ECL | 1484637 | 2316032 | 751161.1 | 64.16 s |
| Policy-Surrogate | 1493636 | 2532428 | 959105.8 | 3.38 s |

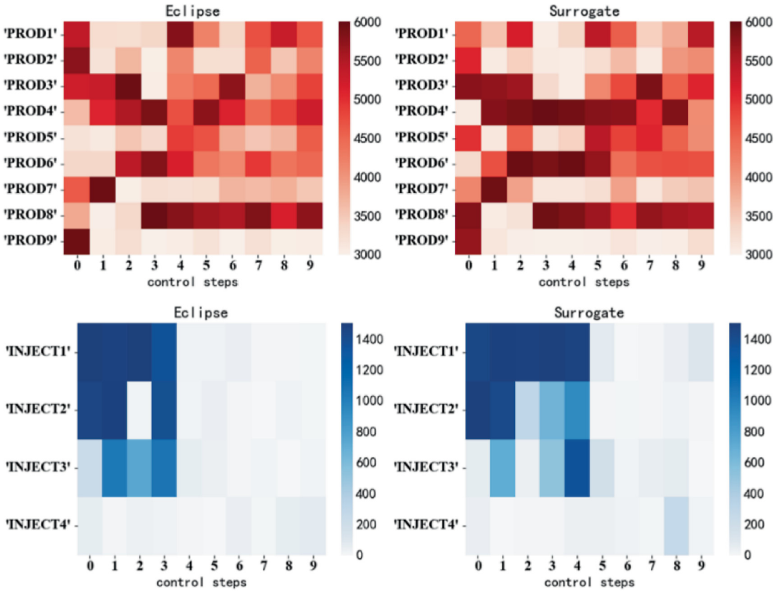


**Fig. 7.** Comparison of injection BHP and production rate (gaussian reservoir model). On the left are the results of interaction with the numerical simulator, and on the right are the results with the help of the state transition model, the bottom hole pressure of the production well is at the top, and the injection well injection rate is at the bottom.

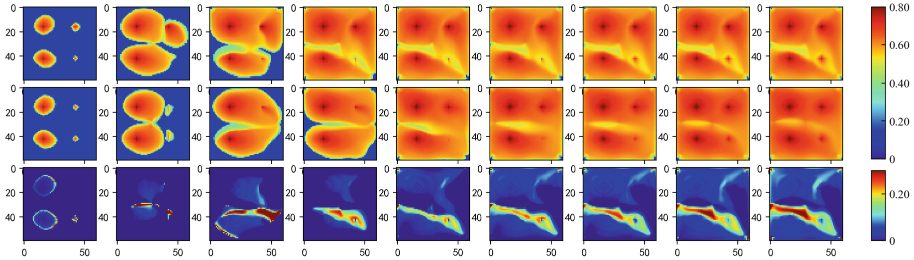


**Fig. 8.** Water saturation of gaussian reservoir model.

## 5 Discussion and Conclusion

In this paper, an off-line reinforcement learning method is proposed to solve the production optimization problem in reservoir which the production optimization problem is defined as a Markov process. Through the continuous learning of agents, the regulation strategy is embedded into the network, and the state transition model can realize offline design of injection-production scheme without numerical simulation.

In future work, the model will be extended to more complicated 3D problems by replacing the method of conv2D with conv3D. Because the state of the three-dimensional model has one more dimension, it is the level information of the model. Furthermore, oil field production needs to consider various practical conditions such as reservoir pressure and water cut, and we aim to incorporate these elements alongside additional nonlinear constraints in forthcoming iterations of the model. Additionally, the joint optimization of well location and production is also one of the future research directions.

**Acknowledgments.** The project is supported by the National Natural Science Foundation of China under Grant 52274057, 52074340 and 51874335, the Major Scientific and Technological Projects of CNPC under Grant ZD2019-183-008, the Major Scientific and Technological Projects of CNOOC under Grant CCL2022RCPS0397RSN, the Science and Technology Support Plan for Youth Innovation of University in Shandong Province under Grant 2019KJH002, 111 Project under Grant B08028.

## References

1. Brouwer D R, Nœvdal G, Jansen J.D, et al.: Improved reservoir management through optimal control and continuous model updating. In: SPE Annual Technical Conference and Exhibition, SPE-90149-MS, OnePetro, Houston (2004)
2. Zhang, K., Zhang, X., Ni, W., et al.: Nonlinear constrained production optimization based on augmented Lagrangian function and stochastic gradient. J. Petrol. Sci. Eng. **146**, 418–431 (2016)
3. Foroud, T., Baradaran, A., Seifi, A.: A comparative evaluation of global search algorithms in black box optimization of oil production: a case study on Brugge field. J. Petrol. Sci. Eng. **167**, 131–151 (2018)
4. Yin, F., Xue, X., Zhang, C., et al.: Multifidelity genetic transfer: an efficient framework for production optimization. SPE J. **26**(04), 1614–1635 (2021)

5. Zhang, K., Zhao, X., Chen, G., et al.: A double-model differential evolution for constrained waterflooding production optimization. J. Petrol. Sci. Eng. **207**, 109059 (2021)
6. Chen, G., Zhang, K., Xue, X., et al.: A radial basis function surrogate model assisted evolutionary algorithm for high-dimensional expensive optimization problems. Appl. Soft Comput. **116**, 108353 (2022)
7. Ma, H., Yu, G., She, Y., et al.: Waterflooding optimization under geological uncertainties by using deep reinforcement learning algorithms. In: SPE Annual Technical Conference and Exhibition, OnePetro, Calgary (2019)
8. Zhang, K., Wang, Z., Chen, G., et al.: Training effective deep reinforcement learning agents for real-time life-cycle production optimization. J. Petrol. Sci. Eng. **208**, 109766 (2022)
9. Wang, Z.Z., et al.: Evolutionary-assisted reinforcement learning for reservoir real-time production optimization under uncertainty. Petrol. Sci. **20**(1), 261–276 (2023)
10. Wang, Z., Zhang, K., Zhang, J., et al.: Deep reinforcement learning and adaptive policy transfer for generalizable well control optimization. J. Petrol. Sci. Eng. **217**, 110868 (2022)
11. Haochen, W., Zhang, K., Chen, N., et al.: Hierarchical optimization of reservoir development strategy based on reinforcement learning. Geoenergy Sci. Eng. **2023**, 211678 (2023)
12. Haarnoja, T., Zhou, A., Abbeel, P., et al.: Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: 35th International Conference on Machine Learning, pp. 1861–1870. PMLR (2018)
13. Zhang, K., Wang, X., Ma, X., et al.: The prediction of reservoir production based proxy model considering spatial data and vector data. J. Petrol. Sci. Eng. **208**, 109694 (2022)
14. Zhang, K., Wang, Y., Li, G., et al.: Prediction of field saturations using a fully convolutional network surrogate. SPE J. **26**(04), 1824–1836 (2021)
15. Ma, X., Zhang, K., Wang, J., et al.: An efficient spatial-temporal convolution recurrent neural network surrogate model for history matching. SPE J. **27**(02), 1160–1175 (2022)
16. Tang, M., Liu, Y., Durlofsky, L.J.: A deep-learning-based surrogate model for data assimilation in dynamic subsurface flow problems. J. Comput. Phys.Comput. Phys. **413**, 109456 (2020)

# Research on Data Governance System of Oil and Gas Field Exploration and Development

Shan-shan Liu(✉) and Xin Liu

Information and Data Center, Sinopec Petroleum Exploration and Production Research Institute,
Beijing 102206, China
lougaosimianfeng@126.com

**Abstract.** Digital transformation is a key factor to improve the efficiency of upstream production companies and accelerate key business decisions in the oil and gas industry. The goal of enterprise digital transformation is to transform traditional business into digital business. The essence of digital business is to process data as a new production factor and build products with data as the main form of existence. With the accelerated development of digital transformation, efficient management and use of massive data has become one of the core challenges in the process of digital transformation of enterprises. Oilfield data governance is an important means to improve oilfield production efficiency and economic benefits, and is also a necessary condition for data driven decision-making. Breaking data islands through effective data governance can promote the application of big data technology. This paper takes the data governance methods and key technologies in the application of oil and gas field exploration and development data as the research object, expounds the data governance system of exploration and development, and gives the solution of data governance and unified data management including data resource inventory, data standards, data quality management, data services and big data operation and maintenance, covering the main contents of the entire life cycle of data governance to solve the data governance problems encountered in the application of big data in the process of exploration, development and production, as well as in the construction of intelligent oil fields.

**Keywords:** Oil and Gas Field Exploration and Development Data · Data Governance · Data Middle Platform

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20-22 September 2023.

This paper was selected for presentation by the IFEDC Committee following review of information contained in an abstract submitted by the author(s). Contents of the paper, as presented, have not been reviewed by the IFEDC Technical Team and are subject to correction by the author(s). The material does not necessarily reflect any position of the IFEDC Technical Committee its members. Papers presented at the Conference are subject to publication review by Professional Team of IFEDC Technical Committee. Electronic reproduction, distribution, or storage of any part of this paper for commercial purposes without the written consent of IFEDC Organizing Committee is prohibited. Permission to reproduce in print is restricted to an abstract of not more than 300 words; illustrations may not be copied. The abstract must contain conspicuous acknowledgment of IFEDC.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 367–379, 2024.
https://doi.org/10.1007/978-981-97-0272-5_32

# 1 Introduction

With the accelerated development of oilfield Digital transformation, the efficient management and use of massive data has become one of the core challenges in the process of enterprise digital transformation. The challenges brought by big data management are reflected in the following aspects: the ability to store and access massive data, the wide range of data sources, diverse data types, and data quality and security.

The upstream oil and gas industry are a complex and data-driven industry, with exponential growth in data volume [1]. Integrating and analyzing information from various data sources has become one of the biggest obstacles to taking quick and decisive action. 60% to 73% of the data within the enterprise is not used for analysis. In order to make the necessary decisions for business management, managers need the latest, complete, and correct data.

The upstream data center gathers various real-time and historical data, including company, drilling, reservoir, production, operations, engineering, cost and economic, business plans, and is provided through various systems and digital formats, including temporary, daily, monthly, quarterly, and annual manual reports.

Research conducted by SPE and global E-P companies has confirmed that data management is a major issue in the oil and gas industry. They have clearly determined that over half of the engineers and Earth scientists' time is spent searching for data and compiling before multidisciplinary analysis begins.

The main business data sources throughout the entire oil lifecycle do not have appropriate integrity, consistency, and data standardization, making it impossible to achieve true integration across all areas of the oil lifecycle. For example, when creating well data in exploration, drilling, reservoir, or production areas, the well data and its main attribute 'Unique Well ID' (UWI) should be the same. Unfortunately, the entire oil company's database may diverge as different UWIs may be found in different regions for the same well. This situation also occurs when examining the name, total depth, and other attributes of the entire organization. Another common situation is that higher-level management requires historical information to make decisions, but this data is not available, or even worse, it no longer exists. The extreme case of using nonintegrated and low-quality data is to use incorrect coordinates to make drilling decisions, which may cause huge losses to the company. Considering these issues and based on existing new technologies, all these scenarios need to be addressed in an integrated manner.

With the continuous growth of data generated by the oil and gas industry, the establishment of a unified sharing platform integrating scientific research, production and management can eliminate information silo, realize the interconnection of exploration and development data, and improve the upstream business operation capability. This makes the need for effective data governance more urgent. Data governance can be used as a method and tool to help enterprises with digital transformation. Effective data governance can quickly enhance the core competitiveness of enterprises. Domestic enterprises in various fields have recognized the importance of data governance and have begun to carry out data governance to assist enterprise management.

Based on the OSDU (Open Subsurface Data Universe) open source data platform, Schlumberger combines business centric data management applications with artificial intelligence platforms to provide intelligent and professional data governance and quality control solutions for energy enterprises.

Saudi Aramco implemented the real-time data management steps of intelligent oilfield in the process of intelligent oilfield data management, including: data standardization, data filtering and compression, data quality control, data quality and transmission key performance indicators, data visualization, and system reliability and availability [2].

The early domestic data governance mostly used data warehouses, which were not associated with businesses and departments, and could not really solve problems. The Digital Oilfield Research Institute of Chang'an University put forward the concept of data governance in 2015, emphasizing the sorting and rectification of data [3]. Sun Min [4] proposed that the oil field data governance should be unified to deal with data islands. Data governance is to solve the problem of multiple databases and inconsistent data standards caused by the construction of multi batch and multi-vendor data. Based on the relevant theoretical research of data governance, the practice needs to be associated with the business department of oilfield enterprises, go deep into the business department to investigate, collect and sort out data, eliminate data inconsistency, introduce standardized data application standards to promote the maximum utilization of data assets in business, management, and enterprise decision-making.

Due to limitations in technological development level and project control, upstream business informatization follows a top-level design divided by business domains and a gradual and rolling construction method that combines group company unified construction with oilfield company self-construction, and basically achieves the goals of digitalization, automation, and networking construction. However, due to the long-term adoption of specialized management and decentralized construction mode divided by business domain, oil field companies and oil service enterprises are faced with the multiple databases, multiple platforms/systems, multiple isolated applications, and the phenomenon of information silo problems, which are difficult to meet the resource sharing needs of enterprise level exploration and development integration, geological engineering integration, and dynamic and static integration. It cannot effectively support the construction of smart oil and gas fields and their future development.

This paper conducts research on data governance system of exploration and development, aiming to establish a complete data management framework, ensure data quality, safety and availability, and provide strong support for organizational decision-making.

## 2 Methodology and Key Technologies of Exploration and Development Data Governance

Oilfield data governance is an important means to improve oilfield production efficiency and economic benefits, and is also a necessary condition for data driven decision-making. The extensive construction of enterprise big data platform has made enterprise digital transformation an inevitable trend. Business and IT jointly support enterprise data governance and digital development, and rapidly enhance the core competitiveness of enterprises.

Data governance is defined as a set of standards, policies, roles and practices to exercise power and control over how to make data decisions, and serve as a guide for people and processes to interact with data. Data governance also aims to solve data management problems and challenges caused by lack of confidence in data.

The International Data Management Association (DAMA) is committed to the research and practice of data management and has published the DAMA Knowledge System and Guidelines for Data Management (DAMA DMBOK), which covers the practical experience of several industry experts worldwide and is of great reference value for researchers in the field of data management. The data governance framework proposed by DAMA (see Fig. 1), covers the ten major functional modules of data governance and is the most systematic data governance framework so far, which can be used as a theoretical guide for enterprises to implement data governance functional modules.

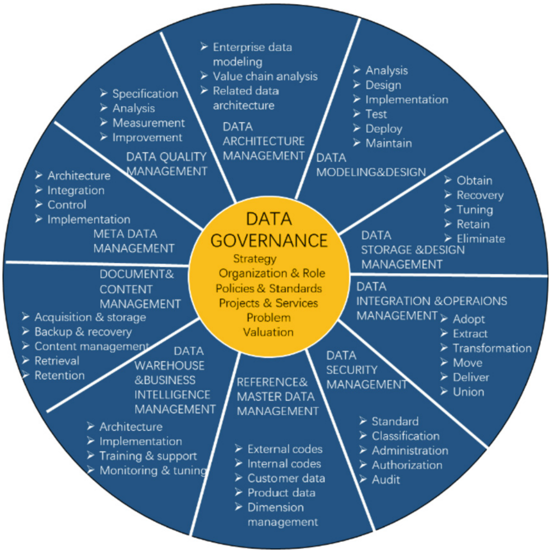


**Fig. 1.** DAMA data governance top 10 function chart.

Professional data management practitioners from various oil and gas operators, software and service companies, with a background of Data Management Association International (DAMA) Data Management Body of Knowledge (DMBOK) version 2 [5], compile and map Professional Petroleum Data Management and practices to the competency areas required by the Certified Petroleum Data Analyst (CPDA) exam certified by PPDM and the Competency Management System (CMS) of the Common Data Access (CDA) organization for oil and gas in the UK. In these exercises, the most consistently selected knowledge areas are data governance, data quality and data security. This paper introduces the data governance system built by oil and gas companies during the construction of the exploration and development data resource center, providing a reference for guiding oil data analysts.

**Key Technologies of Exploration and Development Data Governance.** Data cleaning algorithm: it is used to remove irregular data such as noise, outlier and duplicate values, such as clustering, regression, decision tree, etc.

Data quality analysis algorithm: it is used to detect data quality problems, such as missing values, outlier, inconsistency, incompleteness, data exploration, statistical analysis, etc.

Data standardization algorithm: it is used to unify data from different formats and sources into a consistent standard format.

Data integration algorithm: it is used to integrate data from different sources, such as data integration, data warehouse, data lake, etc.

Data mining algorithm: it is used to discover hidden patterns, patterns, and relationships in data, such as association rules, classification, clustering, etc.

Data security algorithms: it is used to protect the security and privacy of data, such as access control, encryption, desensitization, auditing, etc.

Upstream Data Visualization and Analytics: it is used to present data in a graphical way to facilitate data analysis and decision-making, such as data visualization, report making, and dashboard.

## 3 Data Governance Framework of Exploration and Development

Big data has two values in enterprise decision-making: data sharing, breaking the data barriers and chimney like data architecture between various business departments, and gathering the data resources of various business systems. At the same time, various business systems can also use the data of the data warehouse through data services. The other value is to conduct cross department and cross time joint analysis based on the data of the collected business systems, mining the value of data and assisting various departments in making statistical decisions.

Data governance is a method of organizing data resources, which aims to ensure data quality, security, integrity and availability to maximize data value and credibility. It is a comprehensive management method that manages data within an organization by establishing rules, standards, and processes.

**Process of Implementing Data Governance.** (1) Data management organizations and roles: To break down internal barriers and determine the responsibilities and roles of data management, we should first build a data governance organization and process involving multiple departments to support data management activities and establish a special data governance team. It includes the data governance committee, the data governance team, and the progressive organizational structure of each business department. The steering committee is composed of a leadership team. The data governance team is composed of the chief architect of data governance, the solution architect, the platform technology architect, and industry experts to develop data standards, processes, workflow, and implementation. Any technology provider participating in the company's integrated upstream workflow should follow the framework to support interoperability and avoid creating information silo. Data should be entered at once in an intelligent on-site environment and available throughout the entire asset lifecycle, as well as in various workflows and applications.

(2) Data collection and storage: In oilfield exploration and development, it is necessary to collect and store a large amount of geological, geophysical, engineering and other data. In order to ensure the reliability and consistency of data, it is necessary to establish unified data acquisition and storage standards, and use advanced data management systems, such as data middle ground or data lake.

(3) Data quality management: Data quality is a very important part of oilfield exploration and development. Data quality management includes data cleaning, deduplication, standardization, and other processing to ensure the accuracy, integrity, and consistency of data. It is necessary to use data quality management tools and technologies to achieve data quality management, such as ETL tools, data quality inspection tools, data standardization tools, etc.

(4) Data analysis and modeling: Oilfield exploration and development require the use of a large amount of data analysis and modeling techniques to support geological exploration, reservoir evaluation, production optimization, and other work. It is necessary to establish unified data analysis and modeling standards, and use technologies such as data mining and machine learning to improve the efficiency and accuracy of data analysis and modeling.

(5) Data security management: Data in oilfield exploration and development is very important, and various security measures need to be taken to protect data security, such as access control, encryption, backup, etc.

(6) Data governance organization and process: oilfield exploration and development need to establish a sound data governance organization and process to ensure the smooth implementation of data governance. It is necessary to clarify the responsibilities and roles of data governance and establish corresponding data governance processes, including data collection, processing, storage, analysis, etc.

The exploration and development data center is built according to the "data platform application" model. The data governance project cooperates with the big data platform to make data more valuable by improving data quality.

**Data Governance Framework.** The construction goal of data governance is to support various types of application scenarios by integrating multiple business system data. The construction process is shown in Fig. 2. The big data platform designed and are constructed according to the integrated design concept of unified resource collection, unified data governance, unified data resource management and control, unified data sharing services, and unified big data development application support services, fully access the multi-source heterogeneous big data resources of the exploration and development business system, and realize the integration of the entire process of data resources, from collection, storage and exchange, cleaning and integration, intelligent analysis, data sharing to visual display.

The data warehouse of big data platform is divided into four layers: The first layer of the data warehouse is the paste source layer (ODS), which gathers data from different data sources and accesses them to the data resource pool. Based on HDFS, DBMS, distributed message queue, memory database and other storage technologies, it realizes the storage of massive heterogeneous data (structured data, semi-structured data, unstructured data). Data fusion should fully cover exploration and development business scenarios.

After aggregating data into a resource pool, it is necessary to standardize and clean the data collected by various business systems, including data filtering, deduplication format conversion, and content verification. The cleaned data is put into the standard layer (DWD), and the business flow data is integrated according to the subject domain and business object.

The data of the Theme Layer (DWS) is reorganized, connected, and integrated from a business perspective, summarized into data tables based on dimensions to meet the needs of enterprises for data, and to achieve hierarchical classification and retrieval of exploration and development various theme data assets.

Based on information from various dimensions, API services are generated to support various types of application scenarios. From an application perspective, data aggregation and calculation, as well as encapsulation of calculation logic are carried out to establish

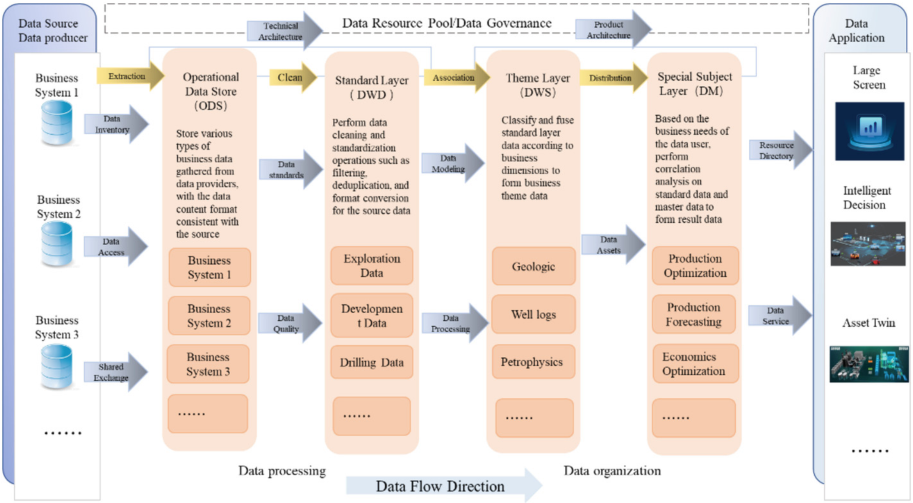
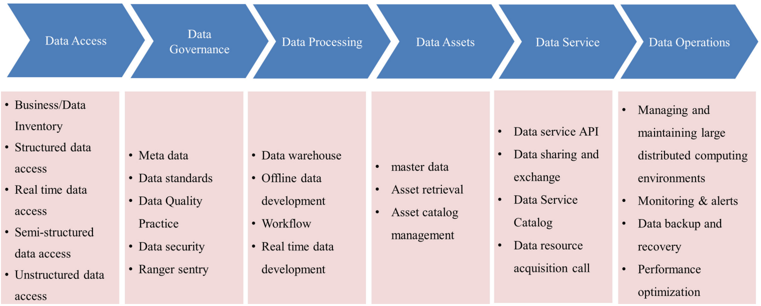


**Fig. 2.** Integrated Architecture and Governance Framework.



| Data Access | Data Governance | Data Processing | Data Assets | Data Service | Data Operations |
|---|---|---|---|---|---|
| • Business/Data Inventory<br>• Structured data access<br>• Real time data access<br>• Semi-structured data access<br>• Unstructured data access | • Meta data<br>• Data standards<br>• Data Quality Practice<br>• Data security<br>• Ranger sentry | • Data warehouse<br>• Offline data development<br>• Workflow<br>• Real time data development | • master data<br>• Asset retrieval<br>• Asset catalog management | • Data service API<br>• Data sharing and exchange<br>• Data Service Catalog<br>• Data resource acquisition call | • Managing and maintaining large distributed computing environments<br>• Monitoring & alerts<br>• Data backup and recovery<br>• Performance optimization |

**Fig. 3.** Data governance content.

corresponding thematic layers (DM) of the application scenarios. The upper level data application can interface with the data service API, which quickly applies data to business scenarios through the form of data services. The data service API interfaces with three common data applications, including data screens, data reports, and intelligent applications.

It can be generally divided into four steps from the data source of each business system to data application: data inventory, aggregation, data control, and data services. First, inventory the exploration and development data from different systems. After the aggregated data is accessed to the platform, a series of data governance measures, such as data standards, metadata and data quality control, are established to form a catalog of exploration and development data resources according to business domains. The data of each subject layer serves each business system through interfaces.

Data governance covers the data source of each business system and data application. The application system supports the data application by calling the data of the data resource pool through services. The core is data capitalization to better support the business. Broad data governance mainly includes data access, data governance, data processing, data assets, data services and data operations (see Fig. 3).

In addition to the above four main processes, it also involves some sub processes, including the main contents of detailed data governance.

## 4 Data Governance Content of Exploration and Development

### 4.1 Upstream Data Inventory

Before data access, it is necessary to conduct data analysis and research on various systems, conduct data inventory, and clarify which systems are relevant to the business domain. What kind of standardization should be achieved in the process of data standardization? It is necessary to construct data standards and apply the technical architecture of data processing. The main task of business research is to find the relationship between business processes and data flows and conduct a three-level analysis.

System level analysis: System usage and importance and system database type should be investigated firstly and the size of big data platforms is determined by the amount of inventory and incremental data.

Table level analysis: Analyze the meaning and improve the information, determine whether to connect and how to connect.

Field level analysis: Improve field information.

Structured data access methods mainly include Sqoop, DataX, Informatica/DataStage Kettle. Figure 4 shows the principle of using Sqoop technology to access data to Hadoop.

The accessed small file data is stored in GlusterFs and large files are stored in DFS. Small images can be stored in Cassandra and other images can be stored in MongoDB, HBase, and structured data is generally stored in Hive.

### 4.2 Data Standards and Integration

The management of master data, metadata and data standards are important part of data governance. Metadata is the data that used to describe data and managed the description

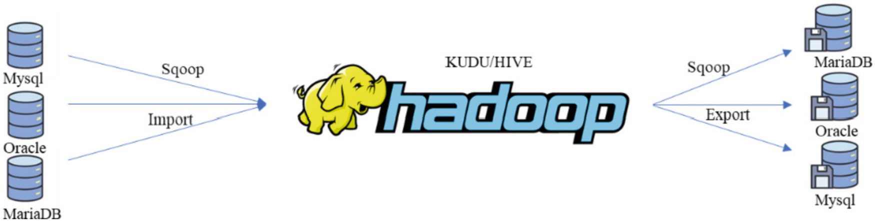


**Fig. 4.** The principle of using Sqoop technology to access data to Hadoop.

information of business data tables during the data warehouse construction process. The description object of master data is the physical data of the core business of the enterprise. Exploration and development master data generally includes oil and gas reservoirs, traps, structural units, development units, wells, station libraries, pipelines, and various instruments and equipment. Due to the differences in various standards of data sources, the issue of inconsistency in the data content aggregated into the resource pool should be addressed before data service sharing or organizational analysis, namely data cleaning. Establishing a unified data standard can achieve the circulation and identification of data assets. Data standards and ontologies provide a robust and reliable data exchange mechanism, enabling data integration and interoperability across multiple databases and systems.

### 4.3 Data Quality Management

The purpose of data quality analysis is that reliable data can make wise decisions. The best recognized practice of data management is to focus on how oilfield operation data supports the most commonly used key data science algorithms in the oil and gas industry, and whether the quality of input data is sufficient to support data science and predictive analysis plans.

The quality of an analytical model depends on the quality of the data it is based on. High quality data has consistent time granularity, sensor resolution, and reading accuracy (see Fig. 5). Only by using high-quality data can we analyze trends and patterns, execute current business operations, make operational decisions, and evaluate future business choices and opportunities. Poor data quality (see Fig. 6) followed by millions of dollars in rework, results in unknown potential value loss due to damaged completion test results, waste of human resources, and nonfactual outputs.

**Fig. 5.** High-quality data used to decisions.

**Fig. 6.** Poor quality, lack of reasonable patterns of data leads to poor decision-making, waste, and limited revenue.

Data quality management involves defining, monitoring, maintaining data integrity, and improving data quality. During the planning and design phases, it is important to define the data elements and quality required for the intended use of each workflow. It is necessary to determine the specific content of data that can be measured or evaluated to understand the quality of the data. For example, DAMA-UK (2013) suggests describing data quality with elements such as completeness, uniqueness, timeliness, effectiveness, accuracy, and consistency [6].

Through on-site research, it was found that there are two reasons for data quality issues: firstly, the slow quality inspection speed. The existing quality inspection methods use manual quality inspection, which takes a long time and there is a lack of quality inspection system and unified management of the quality inspection process, which slows down the quality inspection speed. Secondly, the accuracy of quality inspection is low. The problem of low-quality inspection accuracy is caused by incomplete coverage of quality inspection knowledge, inaccurate knowledge description, slow knowledge update speed, and lack of intelligence [7, 8].

The methods relying on the constraints of the database itself and related application programs to ensure data quality, and controlling data quality through parameter settings still have certain shortcomings that cannot truly achieve automatic audit of data quality.

To improve the accuracy and efficiency of data quality inspection, standardize the process of on-site exploration and development data quality inspection, a detailed analysis of the workflow, business characteristics, and work difficulties of oilfield exploration and development data quality inspection is conducted, the following data quality management improvement methods can be adopted: Establish corresponding quality control measures based on the specific development status of oilfield enterprises and achieve clear control objectives. Develop corresponding business rules for the implementation and design process of data quality.

In the future, it is necessary to continue to improve the proportion of automated and intelligent data quality management, utilize big data algorithms such as neural networks to conduct data quality contro, and design an intelligent quality inspection system for exploration and development data. This may become another direction for data quality control.

## 4.4 Data Resource Directory

The data resource directory is a list of data resources arranged in a certain order after classifying. The data resource directory is used to describe information on what data is available, its structure, the relationships between data, where and how it is stored, and the usage of data. It provides a data "ledger" for data management and data services. By virtualizing physical data through the data resource directory, data is logically integrated. It is an effective means of data resource management and cross departmental collaborative data resource services. The data resource directory implements hierarchical classification of dispersed data resources, which is a logically centralized and physically dispersed "data ledger" that provides unified data resource organization, management, and query services for various users.

The construction of data resource directory is an achievement of data governance, which complements the process of data warehouse construction. The data resource directory can be divided into different dimensions according to different users. When summarizing data from an application perspective during the construction process of the thematic layer, a resource catalog with a business process dimension can be constructed, which helps business personnel quickly locate, search, and use data. According to the exploration and development business process and characteristics, the process of constructing a data resource catalog under the business domain classification dimension is shown in the Fig. 7.

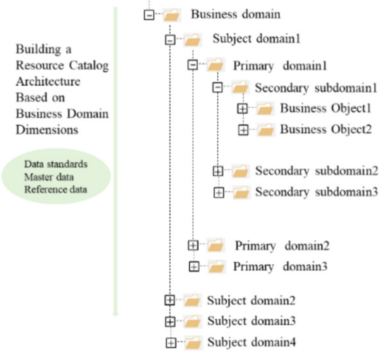


**Fig. 7.** The process of constructing a data resource catalog under the dimension of business domain classification.

According to the different dimensions of the data resource catalog [9], combined with the data resource management needs of each dimension, the primary, secondary, and tertiary classifications of data resources in the data resource catalog vary under different classification dimensions, resulting in corresponding changes. For example, there are three types of data resource directory templates: system classification, business domain classification, and application classification, as shown in the table below (Table 1).

**Table 1.** Data Resource Catalog Template.

| | Primary classification of data resources | Secondary classification of data resources | Three level classification of data resources | Data set | Data item | Unit | Descriptive | Open Property | ...... |
|---|---|---|---|---|---|---|---|---|---|
| System dimension data resource Catalog | System name | Module name | Function name | | | | | | |
| Business theme dimension data resource catalog | Subject domain | Primary domain | Secondary subdomain | | | | | | |
| Application dimension data resource catalog | Application name | Module name | Function name | | | | | | |

## 4.5 Data Services

After the installation and deployment of the big data platform, it can directly support various exploration and development applications, forming a "cloud + end" application ecosystem. During the application development process, data access services are provided based on business system service interfaces, and the data service process is application-Examine-Distribution. Firstly, data users submit applications based on their needs and send data service requests to the server. The approvers examine the data content, scope of use, and sharing method based on the data confidentiality level. Based on the audit results, the data manager obtains business data from the data store after receiving the request, and distributes it after converting it into a specific data format through a configured data API, achieving open application of data sharing. After data distribution, supervision is carried out to ensure the robustness of the system.

The construction of the data center includes a data quality management platform, a data standard platform, a data service platform and a data development platform, etc. The ODS, DWD, DWS, and MD layers of the data warehouse are deployed to the data development platform. The ODS layer includes several thousand data tables, the DWD layer includes one thousand data tables, and the DWS layer includes five hundred data tables. The current focus of the construction of the data resource center is to carry out data governance. Each oil field records data from the past three years and reports on the progress of data rectification work. The data governance project team conducts weekly quality inspections based on quality inspection rules for the data collected from each oil field to the data center. Data quality management platform and a service platform are currently building. The data service platform authorizes API interface applications and provides data services to various system applications.

## 5 Conclusion

Effective data governance is critical for oil and gas operators who use more and more real-time data. Research on the data management system used to support digital energy initiatives will help develop effective data governance practices and help companies better manage their data assets. The framework proposed in this paper can serve as a guide for companies seeking to implement effective data governance practices. By ensuring the accuracy, security, and integrity of data, companies can improve security, efficiency, and reduce operational costs. We look forward to gradually implementing the framework and continuously improving the upstream data base to promote current and future upstream workflows and initiatives, enabling it to move towards higher levels of integrated workflows, visualization, intelligent on-site automation, and advanced analysis.

## References

1. Feblowitz J.: The Big deal about big data in upstream oil and gas. In: Paper & Presentation, IDC Energy Insights (2012)
2. Naser, N.A., Awaji M.A., Aramco, S.: Intelligent field data management: case study. In: SPE/DGS Saudi Arabia Section Technical Symposium and Exhibition Held in Al-Khobar, Saudi Arabia, 15–18 May 2011 (2011)
3. Shaobo, S.: Methods and Technologies Research on Data Governance of Data Application in Oil and Gas Field Exploration and Exploration. Chang'an University, Xi an (2018)
4. Min, S.: Data governance project of intelligent oilfield and its application. China Manag. Informat. **21**(06), 49–50 (2018)
5. Kozman, J.B.: Data management requirements for supporting digital energy initiatives. In: SPE Annual Technical Conference and Exhibition held in Dubai, UAE, 26–28 September 2016 (2016)
6. Lu, Z., Chen, R., Li, Q., Li, Y., Ren, H., Li, G., Yu, S.: Exploration of real-time data quality management methods for drilling engineering in Tarim Oilfield. J. Logging Eng. **31**(03), 118–121 (2020)
7. Zhan, M.: Research and Implementation on Oil Field Development Data Quality Assurance System. Northeast Petroleum University, Da Qing (2016)
8. Xu, L.: Oilfield Development Data Quality Control and Application Research Database and Information System. China University of Petroleum (EastChina), Qing Dao (2014)
9. He, Y.: Research on Construction Technology of Oilfield Data Resource Catalogue. Northeast Petroleum University, Da Qing (2021)

# Carbonate Fracture-Cavity Reservoirs Prediction Technology Based Deep Learning Model

Ning Li[1,2](✉), Ren-bin Gong[2,3], Liang Ren[1,2], Shu-hang Ren[1,2], Jiang- tao Sun[4], Xiao Yu[4], and Chun-ting Gan[4]

[1] Research Institute of Petroleum Exploration and Development, PetroChina, Beijing, China
lining_riped@petrochina.com.cn

[2] Artificial Intelligence Technology R&D Center for Exploration and Development, CNPC, Beijing 100083, China

[3] Technology Consult Center, Research Institute of Petroleum Exploration and Development, PetroChina, Beijing 100083, China

[4] Huawei Technology Co. Ltd., Beijing 100085, China

**Abstract.** Paleozoic carbonate rock is the key field of oil-gas exploration in Tarim Basin. Fracture-cavity reservoirs are often developed in carbonate reservoirs in Tarim Oilfield, a large amount of oil and gas resources are dis-tributed in Paleozoic reservoirs with different burial depths and scales, accurately and quickly identifying these fracture-cavity reservoirs is Significant to the oil and gas exploration, development, and production of the Tarim Oilfield. With the continuous development of artificial intelligence technology, machine learning methods have been widely applied in various scenarios of oil and gas exploration and development, bringing new opportunities for the development of carbonate reservoir prediction technology. Based on drilling, logging and seismic data, this study comprehensively analyzes the structure, rock and physical properties of carbonate reservoirs in the study area, exploring the main controlling factors of carbonate reservoirs. On this basis, a sample set corresponding to the fracture-cavity reservoirs in the study area was constructed, by using machine learning methods, a prediction model for carbonate rock fracture-cavity reservoirs has been established, which can intelligently predict carbonate rock fracture and cave reservoirs in the research area. The trained model of carbonate rock fracture and cave reservoir prediction can quickly and accurately identify fracture and cave reservoirs on post stack seismic data. This study demonstrates that methods are based on machine learning can quickly and efficiently predict carbonate rock fracture-cavity reservoirs.

**Keywords:** artificial intelligence · oil and gas exploration and development · carbonate rocks · reservoirs · deep learning

## 1 Introduction

The concept of artificial intelligence (AI) can be traced back to the early 1950s. According to records, the German mathematician Alan Turing proposed the "Turing test" in his paper "Computing Machinery and Intelligence," which is also the foundation of AI. In

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 380–392, 2024.
https://doi.org/10.1007/978-981-97-0272-5_33

addition, many people have put forward the concept and ideas of AI in other fields, including automata, neurons, neural net-works, etc.

Early AI technologies [2] were based on symbolic logic and expert systems. At this stage, researchers mainly transformed the knowledge of human thinking processes into logical rules that computers could under-stand, in order to achieve AI. However, due to the lack of sufficient computer resources, the development of AI technology at this stage was very limited and unable to handle complex problems.

From the 1980s to the early 2010s, "neural networks" became a focus of research. Neural networks are an algorithm inspired by biology, which allows computers to continuously improve their recognition and judgment abilities through learning. At this stage, AI technology mainly focused on achieving speech recognition and computer vision. Due to the lack of sufficient training data and resources, the application of neural network technology still has significant limitations.

With the rapid development of data collection technology and computer capabilities, deep neural networks have been widely applied to areas such as speech recognition, image recognition, natural language processing, etc. The emergence of deep learning technology has enabled AI to make rapid progress and breakthroughs in these fields.

As artificial intelligence technology develops rapidly, various industries are exploring how to apply AI to their own business. In the oil industry, AI technology has been widely used and achieved significant results. Currently, AI is mainly applied in exploration block evaluation, seismic data processing and interpretation, reservoir evaluation, and reservoir development optimization in the process of oil exploration and development. The PetroChina cognitive computing platform uses AI technology to achieve automatic processing and interpretation of seismic data, intelligent recognition of oil and gas zones and intelligent prediction of production, greatly improving efficiency and accuracy.

Carbonate rock reservoir and Clastic rock reservoir are the most important reservoir types in oil exploration and development. Around the late 19th and early 20th centuries, the first carbonate reservoir was discovered in North Dakota, USA, called Eagle Cape Oilfield. Since then, a continuous wave of oil and gas extraction has occurred in the United States and globally. The largest carbonate field is located in Jeddah, Saudi Arabia, with an estimated output of more than 75 billion barrels of oil. Now, the annual coalescence rate of carbonate rocks has accounted for 60% of the total production, and carbonate reservoirs have become one of the main oil and gas reservoirs.

With the continuous development of oil exploration technology and geological exploration technology, people's understanding of carbonate rock reservoirs is gradually deepening. In the early stage of carbonate rock reservoir prediction, geologists mainly relied on their experience to judge the existence of reservoirs. For example, they studied information such as fractures, caves, lithology in the reservoir, compared it with known oil and gas reserves to predict new ones. This method mainly relies on geologists' experience and lacks high-precision pre-diction capabilities.

In order to improve the accuracy and reliability of carbonate rock reservoir prediction, some prediction methods based on physical and mathematical models have gradually emerged. By using logging and inversion methods, the parameters such as reservoir lithology, porosity, etc. are transformed into visual unit models to further determine the existence of the reservoir. This method is more scientific and accurate than traditional methods but requires a lot of manual work and time.

With the development of AI technology, the prediction of carbonate rock reservoirs will also enter a new stage. AI technology can automatically analyze a large amount of geological exploration data, extract effective reservoir features, and use intelligent prediction models to make carbonate rock reservoir predictions through intelligent analysis. Therefore, based on our actual research area, we characterized the samples of cracks at different scales in seismic data, trained neural networks to achieve efficient automatic identification. In this study, we applied convolutional neural network methods to achieve intelligent prediction of carbonate fractured-vuggy reservoir.

## 2 Research Area Overview

The Tarim Basin [1] is located between the Tianshan Mountains, the Kunlun Mountains, and the Altai Mountains, and is a large basin composed of Mesozoic Cenozoic foreland basins and Paleozoic basin. The study area is located in the northern part of the Tarim Basin, which is part of Lunnan low swell. It is a long-term developed giant ancient uplift formed on the pre Sinian metamorphic basement. It develops marine sedimentary strata from Sinian to Devonian, marine continental sedimentary strata from Carboniferous to Permian, and terrestrial sedimentary strata from Triassic to Quaternary. Tabei area is the main distribution area of Carbonate rock reservoirs in Tarim Basin, where the main oil-bearing series are Lianglitage Formation and Yijianfang Formation. The Lianglida formation is generally sedimentary in the water-infiltration-high-position system domain against the background of the platform. The lower strata are mainly composed of mixed concretions of light green-gray, grayish white, and brown limestone, which are products of water-infiltration system domain sedimentation. The upper pure granular limestone with an unconformably developed karst cave and cracks is a good reservoir layer. Integrated contact between Yijianfang formation and the overlying Tumuxiuke formation. The lithology is mainly composed of light brown gray and grayish brown bright crystalline sandstone limestone, bright crystalline oolitic limestone, and bright crystalline algal sandstone limestone. The biological particles are mainly composed of blue-green algae and their debris, with biological burrows and disturbance structures, making it a good reservoir layer (Fig. 1).

Due to multiple periods of karst activity, there are multiple karst zones and multiple stages of fault activity in the Tarim North Block [4]. The fault distance is small, activity is strong, and the fracture plane combination is complex. The Tarim North Block mainly develops strike slip faults, which began in the Early Caledonian period and formed during the Late Caledonian Early Hercynian period. In the late stage, tensile faults were superimposed and developed. The strike slip [6] faults not only serve as the oil and gas transportation channels for the Tarim Ordovician carbonate reservoir but also are the main areas of karst reservoir development and transformation. Oil source faults control the reservoir characteristics of the study area, determining the development degree of oil and gas reservoirs.

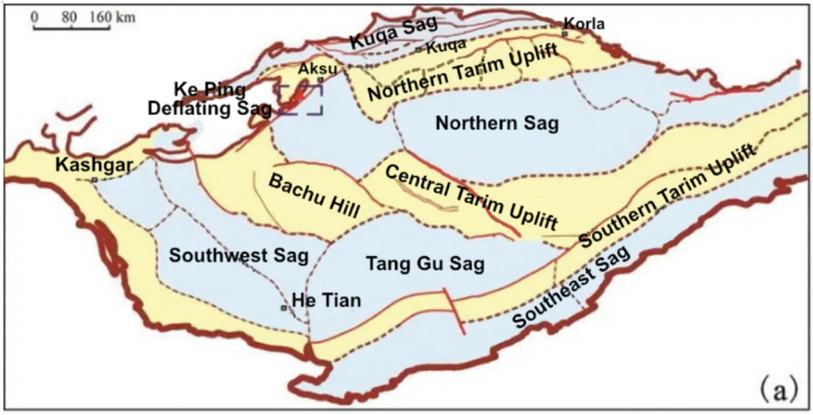


**Fig. 1.** Structure of Tarim Basin

The enrichment pattern of reservoirs in the study area is not only controlled by faults [4], but also influenced by karst processes. From north to south, both fault and karst activities control the differences in reservoir distribution. Reservoirs in fault-controlled areas are mainly developed by fault expansion karst and are the main areas of fault expansion karst development zone [5]. Karst reservoirs have a large scale in vertical development and linear distribution along the fault zone. Based on the geological background of fractured-vuggy reservoir [3] development, the karst reservoirs in the study area can be divided into four types of controlled reservoir development: buried-hills weathering type, underground river in bedding zone, interlayer type in platform edge, and fault-controlled type in karst development zone (Fig. 2).

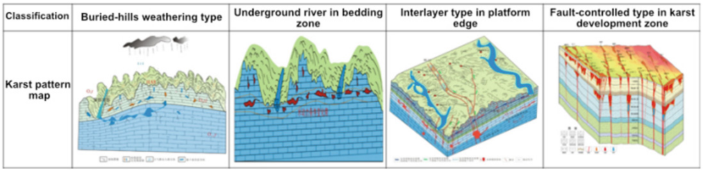

Classification
Buried-hills weathering type
Underground river in bedding zone
Interlayer type in platform edge
Fault-controlled type in karst development zone
Karst pattern map

**Fig. 2.** Division of karst reservoirs in the research area

Through the combination of static and dynamic re-search, the reservoir types in the study block can be classified as cave type [3] and fractured-vuggy type. Overall, cave type reservoirs dominate. From the analysis of the distribution of reservoir types in the region, buried-hills zone, bedding zone, fault-controlled zone are mostly cave type reservoirs, while the platform edge zone is mainly composed of fractured-vuggy reservoir.

## 3 Research on Artificial Intelligence Methods

Machine learning is a process of solving optimization problems, utilizing certain mathematical methods to learn from input samples, building corresponding models, and continuously adjusting and optimizing model parameters to increase model fitting ability. In this study, support vector machine (SVM), random forest, and convolutional neural network (CNN) were studied, and the three methods were compared and analyzed. At the same time, model parameters were continuously improved and optimized, laying the foundation for future research on carbonate rock fracture-void reservoir prediction.

SVM [8] is a classic binary classification algorithm, of-ten used for text and hypertext classification, face recognition, and other problems. Given a training set $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}, y_i \in \{-1, 1\}$, SVM learns the basic idea of finding a hyperplane to separate different sample categories based on the training set D. For linearly separable data sets, SVM can be solved using convex optimization; if the data set is nonlinearly separable, kernel functions are used to map the training set to higher-dimensional space, and then the hyperplane that separates the data set linearly in higher-dimensional space is solved (Fig. 3).

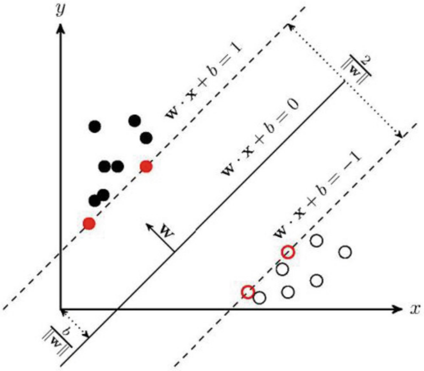


**Fig. 3.** Support vector machine algorithm functions

Random Forest [9] is an algorithm of ensemble learning. Ensemble learning completes prediction tasks by building and combining multiple learners. Random Forest integrates the prediction results of multiple decision trees through the method of ensemble learning, where each decision tree is equivalent to a learner. Assuming there are N learners, the output will have N results. The Random Forest algorithm selects the result with the highest vote count as the final output through voting. Since it can process both continuous and categorical data simultaneously, Random Forest can not only be used for regression tasks but also for classification work (Fig. 4).

Convolutional Neural Network (CNN) [7] is a neural network with deep structure centered on convolutional layers, activation layers, and pooling layers. The role of convolutional layers is to extract local features from the input layer through convolution operations, and feature maps are the extracted features of time series or images after

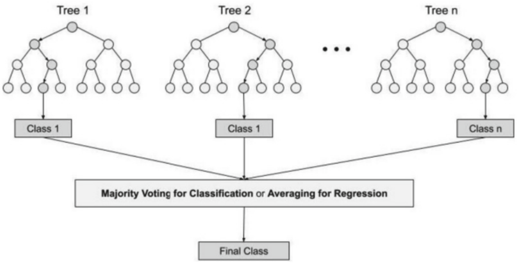

Tree 1
Tree 2
Tree n
Class 1
Class 1
Class n
Majority Voting for Classification or Averaging for Regression
Final Class

**Fig. 4.** Random Forest principle

convolution; the role of pooling layers is to select features extracted from convolutional layers, reduce feature numbers to decrease parameter numbers and avoid overfitting; activation layers mainly de-linearize the in-put of convolutional layers through activation functions, thereby improving the network's learning ability for the training set. Finally, all extracted feature information will be summarized by a fully connected layer, and then classified based on the summarized information (Fig. 5).

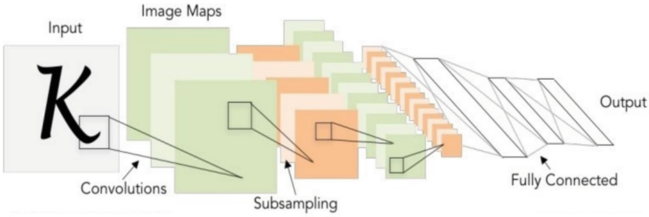


**Fig. 5.** Convolutional Neural Network

Through the above research, it is found that these three artificial intelligence learning methods have their own characteristics: support vector machine (SVM) hyperplanes are determined by a small number of support vectors, which can well eliminate the impact of noise on the results, therefore SVM algorithm has strong robustness; however, the SVM algorithm depends on quadratic programming to solve, which will consume a lot of machine memory and operation time, so it is only suitable for processing small sample learning problems. The Random Forest algorithm combines the results of multiple decision trees, greatly improving the model's generalization ability; however, the Random Forest algorithm is very sensitive to training set noise, so there is still the problem of overfitting. Convolutional neural network (CNN) has strong feature learning ability on the input layer, and because of the flexible structure of CNN, pooling layers or unification layer can be added to improve the model's generalization ability and reduce the risk of overfitting by reducing the number of model parameters or changing the parameter distribution of the model. Moreover, convolution operations have mature algorithms and are easy to parallelize, which can greatly improve the calculation speed and handle large input sets.

Seismic data have the characteristics of large data volume and high noise level, and due to the fact that fractured-vuggy samples are much fewer than non-faulted samples, it requires algorithms with strong feature learning ability. After comparing the above three algorithms, it is concluded that CNN is suitable for processing large data sets and has strong fitting and generalization abilities. Therefore, this article chooses CNN method for pre-diction of carbonate fractured-vuggy reservoir.

# 4 Research on Prediction of Carbonate Fractured-Vuggy Reservoirs Based on Artificial Intelligence

## 4.1 Sample Selection

The research area has a large number of wells, abundant logging data, and good seismic data. In this study, the identification of fractured-vuggy reservoir is mainly carried out using combined seismic information, drilling data, and logging interpretation data.

The seismic reflections of carbonate rock in the re-search area can be classified into three types of fractured-vuggy reservoir responses: string-like reflection, patchy reflection, and chaotic reflection. These are also the main objectives of exploration and drilling in the re-search area. String-like reflection reflects the overall seis-mic characteristics of large caverns, fault clusters, and dense fracture zones. The seismic string-like reflection is a response characteristic of karst reservoirs on seismic planes. The string-like reflection is clear on conventional seismic planes, but the boundary of the string is difficult to identify. Patchy reflection refers to strong energy seis-mic reflections with patchy features, often associated with fractures. Chaotic reflection requires the ability to reflect changes in seismic reflection amplitude and describe changes in local strata formation (Fig. 6).

Combining the above mentioned characteristics of seismic reflection types, this study used geological statis-tical inversion methods and processes. By applying rock physical analysis techniques and geological statistical inversion techniques, high-resolution impedance in this area was inverted. Then, through well logging multi-well consistency interpretation, accurate well logging interpretation results and seismic longitudinal wave impedance relationship models were obtained. For the research area, the main impedance intervals of fractured-vuggy reservoir are $< 166000$ (Fig. 7).

Fractured-vuggy reservoir identification is based on drilling information, well logging responses, and imaging data. Drilling and logging results for carbonate fractured-vuggy reservoir show emptying and leakage. Drilling loss points are the most direct reflection of reservoir development. Based on the magnitude of leakage, the reservoir type at this loss point can be defined. The loss-emptying sections cannot obtain well logging curves, so they are all considered to be high-quality reservoirs. Common well logging response characteristics of carbonate fractured-vuggy reservoir are low natural gamma value, low resistivity, and obvious increase in acoustic wave, corresponding to effective porosity $\geq 1.8\%$ in well logging interpretation.

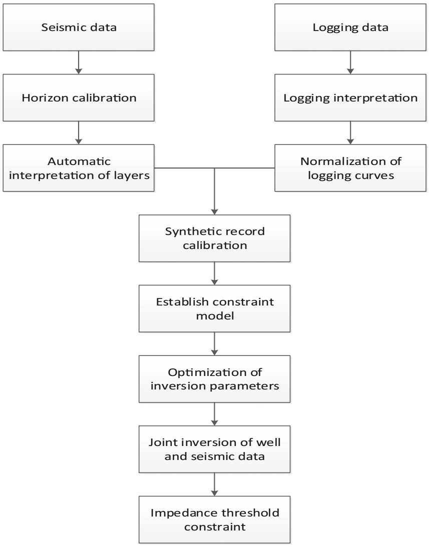
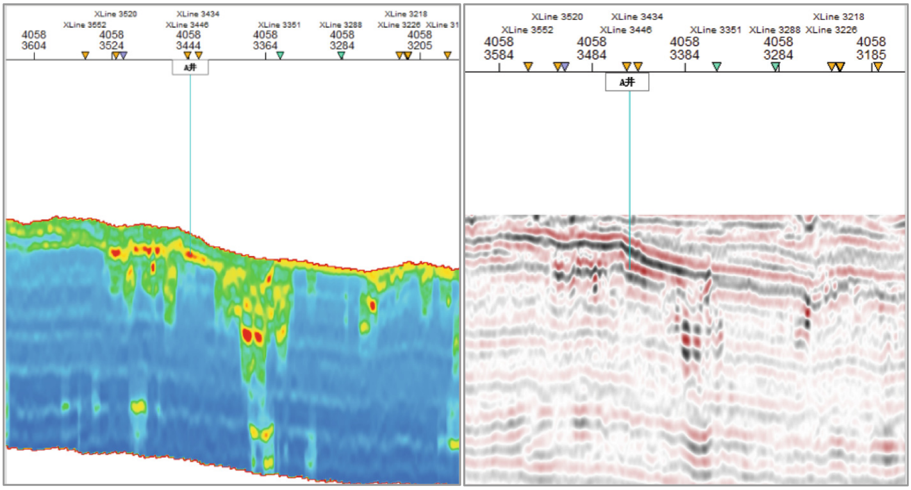

Seismic data
Logging data
Horizon calibration
Logging interpretation
Automatic interpretation of layers
Normalization of logging curves
Synthetic record calibration
Establish constraint model
Optimization of inversion parameters
Joint inversion of well and seismic data
Impedance threshold constraint

**Fig. 6.** Reservoir inversion process



(a) Impedance inversion profile (b) Seismic profile

**Fig. 7.** Sample calibration

The type, thickness, and velocity of the ground rock vary greatly vertically and horizontally, and there is no good refraction interface. Because the thickness and velocity of the ground surface vary greatly along the lateral low-velocity zone, there are even rock formations exposed in some places. In addition, the seismic wave energy absorption and attenuation are serious, there is no low-velocity layer, and the excitation and reception effects are poor (Fig. 8).

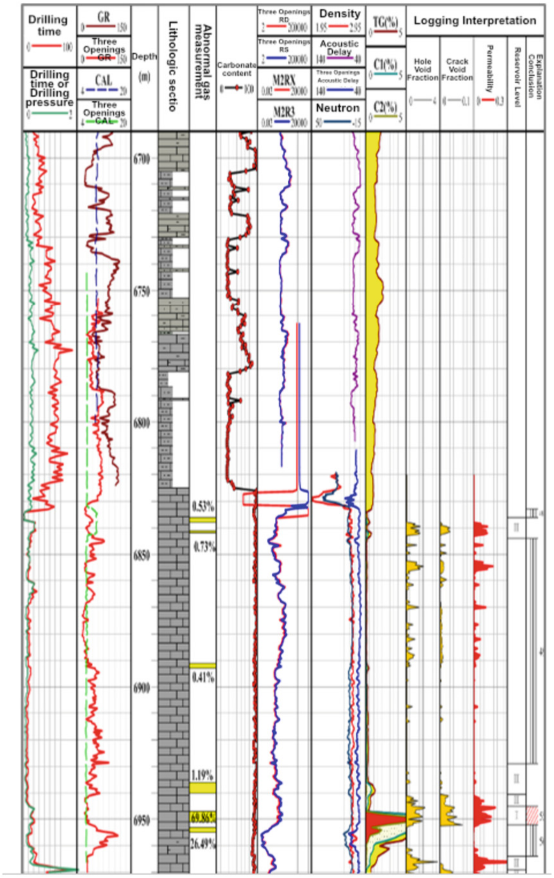


**Fig. 8.** Well logging response characteristics of carbonate reservoir

## 4.2 Data Processing

Our training set contains 1,632,750 pairs of one-dimensional raw amplitude and instantaneous amplitude extracted manually. Each pair of amplitude data corresponds to a (inline, crossline) coordinate. The length of the one-dimensional amplitude data is 365 points with a sampling interval of 5 s. As the oil-water trend surface is a strong reflection surface for seismic waves, the amplitude has a strong response when seismic waves pass through the trend surface. To filter out the interference of amplitude data near the trend surface on the model training process, we started sampling 16 points below the trend surface and sampled a total of 100 points. We used the time window method to process the sampling point set and cut it into multiple time windows of size 6 points, with each time window sliding by one point. If the size of a fracture body in the time window exceeds 80%, it is marked as a positive sample; if there are fracture bodies in the time window but their proportion is less than 80%, they are discarded; if there are no fracture bodies in the time window, they are marked as negative samples. Since the number of negative samples is much larger than positive samples, to balance the sample distribution, we deleted unnecessary negative samples to make the number of negative samples equal to that of positive samples. After screening, the total number of samples is 522,222, and we divide them into 90% for training and 10% for testing. To meet the input requirements of convolutional neural networks, we converted the label vector into the one-hot format. To prevent overfitting caused by erroneous labels, we also performed label smoothing on the label vector, with the parameter setting of 0.3.

### 4.3 Establishment and Training of Recognition Models

We choose the classic ResNet18 [7] structure for convolutional neural networks. The core idea of ResNet is to add straight connections to the non-linear convolutional layers to ensure that the gradients at deeper levels can be propagated back to the shallower levels, avoiding the problem of gradient vanishing when the number of convolutional neural network layers increases. The complex network structure of ResNet18 ensures that the convolutional neural network has sufficient fitting ability. Moreover, ResNet18 adds batch normalization layers, which adjust the weight distribution of each neuron to a normal distribution, enabling the neural network to have sufficient generalization ability. The core layer of ResNet18 is $\begin{bmatrix} 3 * 3\ 64 \\ 3 * 3\ 64 \end{bmatrix}$, which consists of two convolutional layers with 64 $3 \times 3$ convolution kernels stacked on top of each other. ResNet18 mainly consists of eight core layers, an $7 \times 7$ convolution layer, and a fully connected layer, with a total number of layers of $8 * 2 + 1 + 1 = 18$.

For model optimizer, we use Adam optimizer, where the update formula for model parameters is: $\theta_t = \theta_{t-1} - \nabla \mathrm{L}(\theta_{t-1})$, where θ is the model parameters, η is the learning rate set to 0.004, and L is the loss function. We use mini-batch stochastic gradient descent optimization algorithm with a batch size of 1024 and a total number of iterations set to 2 (Fig. 9).

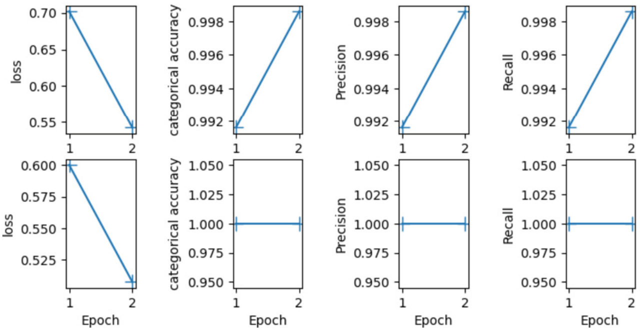


**Fig. 9.** Results of model evaluation

### 4.4 Prediction Effect

By using convolutional neural network methods to learn the sample set, a corresponding carbonate fractured-vuggy reservoir prediction model is established. This study predicts the original seismic data using a learned fractured-vuggy reservoir prediction model. The research area of this study is a slope structure that plunges towards the southwest with developed strike-slip faults (Fig. 10).

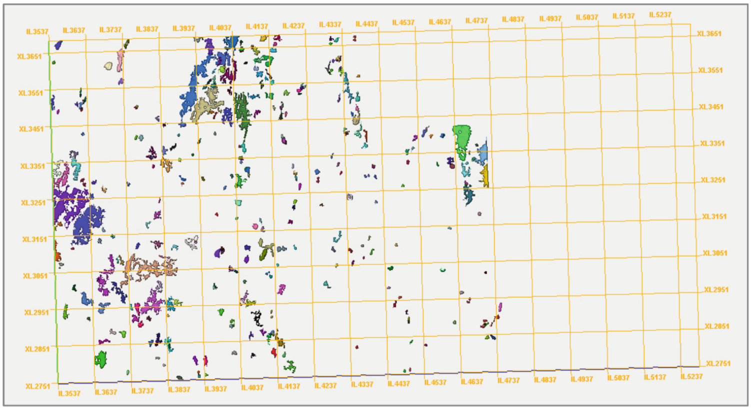


**Fig. 10.** Reservoir prediction results based on artificial intelligence

The high value areas of the prediction results are mainly distributed in Slope structure area with fault developed. In this study uniformly selected 9 wells as validation wells in the research area, the real drilling data and logging interpretation results were compared with the prediction results. The prediction results are shown in the table as follows, including: depth, gamma, AC (acoustic wave), target result, prediction result (Table 1).

**Table 1.** Reservoir prediction results based on artificial intelligence

| Number | Depth | Gamma | AC | Target result | predicted result |
|---|---|---|---|---|---|
| 1 | 6945 | 30 | 52 | 1 | 1 |
| 2 | 6957 | 27 | 50 | 1 | 1 |
| 3 | 6922.5 | 97 | 98 | 1 | 1 |
| 4 | 7120.5 | 10 | 50 | 1 | 1 |
| 5 | 7007 | 22 | 56 | 1 | 1 |
| 6 | 7064 | 13 | 50 | 1 | 1 |
| 7 | 6709 | 59 | 52 | 1 | 0 |
| 8 | 6891 | 14 | 50 | 1 | 1 |
| 9 | 7015 | 15 | 49 | 1 | 1 |

After statistical analysis, the convolution neural network results predicted the 246 fracture body with an accuracy rate of 88%. Through the above studies, it is found that the convolutional neural network method established a good recognition ability for the carbonate fractured-vuggy reservoir prediction model. The 1 samples in the table correspond to the fracture bodies in Fig. 11, and the identification results are consistent with the real drilling results.

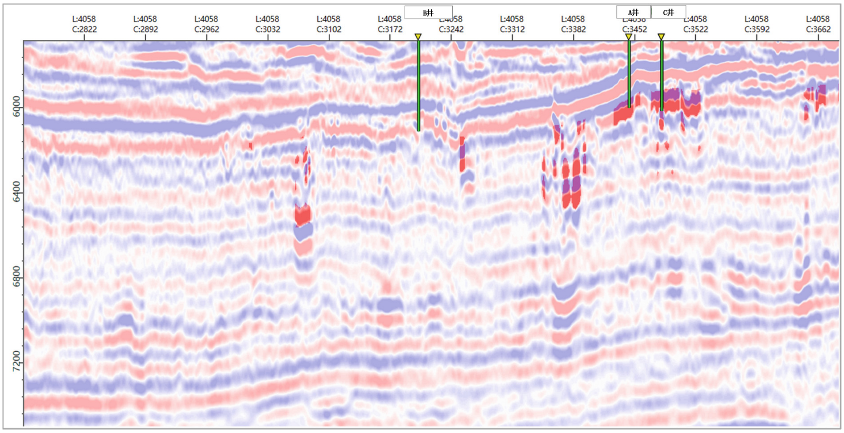


**Fig. 11.** Inline 4058 Seismic profile

## 5 Conclusion

A comparative study of support vector machine, random forest, and convolutional neural network methods was conducted for the identification and prediction of carbonate rock fracture reservoirs. The convolutional neural network algorithm with better prediction performance was selected to perform carbonate rock fracture reservoir prediction in the research area. The following conclusions were obtained:

(1) Support vector machines will consume a large amount of machine memory and computation time, the Random forest algorithm still has the problem of over fitting. Convolutional neural network has strong feature learning ability, which can improve the generalization ability of the model and reduce the risk of over fitting. Therefore, Convolutional neural network is suitable for processing large data sets and has strong fitting and generalization ability.

(2) Based on the convolution neural network method, the prediction results for the reservoirs in the northern part of Tarim Basin show that the reservoirs near the fault are more developed, which is consistent with the geological understanding and single well statistics, and more accurately reflects the development characteristics of real carbonate rock fracture reservoirs.

**Acknowledgments.** The work is supported by the following projects: ①CNPC Project “Research on Intelligent Interpretation Technology and Knowledge Graph of Oil and Gas Exploration” (Number: 2021DJ7003); ②CNPC RIPED project “Research on integrated mathematical model of fracture network and software module development” (Number: YGJ2019-07-04). ③CNPC major scientific and technological project of upstream key core technology "Volume fracturing optimization design software" subproject "EDFM-based Component reservoir simulation software" (Number: 2020B-4118). ④CNPC RIPED project “Joining 5 industrial alliance organizations such as the University of Calgary for exchange and cooperation research” subproject “Joining the MMRD Consortium at Colorado School of Mines” (Number: 2021DQ0105–03). And the author would like to express the sincere gratitude to the following researchers and experts, including Xiao-guo Wan, Bao-zhu Guan, Tao Su, Ji-qing Wang, Xing-yin Zhang and Jun-jie Liu, and.

## References

1. Cao, Y., Zhang, Y., Shen, A., et al.: Carbonate reservoir formation and hydrocarbon accumulation of Ordovician in Gucheng area, Tarim Basin. Marine Origin Petrol. Geol. **25**(4), 303–311 (2020)
2. Ben-Awuah, J., Padmanabhan, E.: An enhanced approach to predict permeability in reservoir sandstones using artificial neural networks (ANN). Arabian J. Geosci. **10**(7), 173 (2017). https://doi.org/10.1007/s12517-017-2955-7
3. Li, H., Dou, Z., Wang, S., et al.: Seismic multi attributes recognition for carbonate fractured-vuggy reservoirs with “weak reflection” characteristics. Geophys. Prospecting Petrol. **53**(6), 713–719 (2014)
4. Ni, X., Zhang, L., Shen, A., et al.: Paleo-karstification types karstification periods and superimposition relationship of Ordovician carbonates in northern Tarim Basin. Geology of China **36**(6), 1312–1321 (2009)
5. Ning, C., Sun, L., Hu, S., et al.: Karst types and characteristics of the Ordovician fracture-cavity type carbonate reservoirs in Halahatang oilfield. Tarim Basin. Acta Petrolei Sinica **42**(1), 15–32 (2021)
6. Wan, X.-G., Guang-hui, W., Xie, E., et al.: Seismic prediction of fault damage zones in carbonates in Halahatang area, Tarim Basin. Oil Gas Geol. **37**(5), 786–791 (2016)
7. Chang, D.-K., Yong, X.-S., Wang, Y.-H., et al.: Fault identification method for seismic data based on deep convolutional neural network. Oil Geophys. Prospect. **56**(1), 1–8 (2021)
8. Li, Y.-F., Cheng, J.-Y., Wang, C.: Seismic attribute optimization and coalbed methane prediction based on support vector machine. Coal Geol. Expl. **40**(6), 75–78 (2012)
9. Zhi-peng, Q., Wang, F.-F., Zhang, Y.-Y., et al.: Seismic multi-attribute sand body thickness prediction based on association rules and random forest. Bull. Geol. Sci. Technol. **40**(3), 211–218 (2021)

# Construction of Fracturing Knowledge Graph and Fracturing Plan Optimization

Xia Lin[1,2], Chao Xu[1,2](✉), Lan Mi[1,2], Zong-shang Liu[1,2], Chong Xiang[1,2], and Li-xia Liu[1,2]

[1] PetroChina Research Institute of Petroleum Exploration and Development, Beijing, China
xuchao1988@petrochina.com.cn

[2] Artificial Intelligence Technology R&D Center for Exploration and Development, CNPC, Beijing, China

**Abstract.** As an efficient and intelligent means of knowledge organization, knowledge graph has become the core force driving the development of artificial intelligence. Hydraulic fracturing is an important measure for increasing production and injection in oil and gas fields, with complex design processes and numerous influencing factors. In order to achieve rapid and accurate optimization of fracturing plan, this paper proposes a method for optimizing fracturing plan based on knowledge graph. By combing the system of fracturing domain knowledge, fracturing knowledge graph is constructed. Extracting characteristic parameters describing the geological engineering double sweet spot in multiple dimensions and multiple scales, and showing the characteristic parameter-related entities, relationships, and attributes as vectors via graph embedding technique. Integrate expert knowledge with artificial intelligence to build a fracturing effect prediction model and optimize the fracturing plan. In this study, more than 500 fracturing oil wells in a tight sandstone block are taken as objects to build a knowledge graph. Based on well test and production test data and historical production, this study predicts the fracturing stimulation effect and optimizes the fracturing engineering parameters. The calculation results indicate that factors such as reservoir thickness, oil saturation, number of fracture clusters, and half length of fractures have a significant impact on the fracturing effect. The coincidence rate between the predicted capacity of production and the actual capacity of production is over 91%, and the efficiency of fracturing plan design is increased by more than 20 times. The research results can provide scientific basis for predicting fracturing effects and optimizing fracturing engineering parameters, greatly improving

This paper was selected for presentation by the IFEDC Committee following review of information contained in an abstract submitted by the author(s). Contents of the paper, as presented, have not been reviewed by the IFEDC Technical Team and are subject to correction by the author(s). The material does not necessarily reflect any position of the IFEDC Technical Committee its members. Papers pre-sented at the Conference are subject to publication review by Professional Team of IFEDC Technical Committee. Electronic reproduction, distribution, or storage of any part of this paper for commercial purposes without the written consent of IFEDC Organizing Committee is prohibited. Permission to reproduce in print is restricted to an abstract of not more than 300 words; illustrations may not be copied. The abstract must contain conspicuous acknowledgment of IFEDC. Contact email: paper@ifedc.org.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 393–406, 2024.
https://doi.org/10.1007/978-981-97-0272-5_34

the efficiency and quality of fracturing design, and improving the success rate of fracturing construction.

**Keywords:** hydraulic fracturing · knowledge system · knowledge graph · graph embedding technique · prediction of fracturing effect · optimization of fracturing plan

## 1 Introduction

As an important branch of artificial intelligence technology, knowledge graph has powerful semantic expression, storage and reasoning functions, and has become a research hotspot in recent years. The knowledge graph has achieved a qualitative leap in knowledge representation and reasoning. Its core technology is to transform complex data information into a visual map, which solves the problem that the required data cannot be obtained quickly and accurately due to the excessive amount of information [1, 2].

The field of fracturing involves multiple stages of oil and gas exploration and production, with a wide range of related majors and disciplines. The data on geology, oil and gas reservoirs, drilling, logging, oil testing and production, fracturing, and oil and gas production are both interrelated and independent. Therefore, this field requires interdisciplinary knowledge and has unique domain complexity. The current problems faced by the optimization of fracturing schemes include: ① cumbersome basic data sorting in the early stage, large workload, and low efficiency; ② Unconventional reservoirs have characteristics such as low porosity, low permeability, and strong heterogeneity, greatly increasing the difficulty of fracturing design and construction; ③ The design of fracturing schemes is highly professional and heavily relies on expert experience.

The scientific utilization of these multi-source, heterogeneous, and massive data is of great significance, but its technical challenges are very significant, and there is an urgent need to develop new technologies and methods to efficiently tap into their enormous value. Knowledge graph has strong advantages in big data analysis, intelligent recommendation and interpretable AI [3–5]. The retrieval method based on knowledge graph is based on entities and retrieves through a large number of relationships between entities, and returns the attributes and attribute values of entities to users. Knowledge graph can fuse structured, semi-structured and unstructured data, and the knowledge storage mode of graphical model is conducive to mining the internal association of knowledge [6]. The knowledge graph can describe different concepts, entities and their relationships in the real world in a structured way [7]. It can express the knowledge in the fracturing field into a more acceptable form through the knowledge graph, sort out the system architecture of the knowledge in the fracturing field, and give better play to the value of relevant data in the fracturing field.

## 2 Construction of Knowledge Graph in Fracturing Field

The construction of knowledge graph mainly includes five parts: ontology construction, knowledge acquisition, knowledge fusion, knowledge storage, and knowledge application [8]. The construction of the knowledge graph in the fracturing field should follow the principle of business driven design. According to the characteristics of different types of oil and gas reservoirs, it should integrate a large number of multi-source structured data related to fracturing design and construction, such as logging, logging, oil testing, drilling, fracturing and production performance, as well as unstructured data such as research reports, literature and multimedia. It is classified by knowledge system, built by ontology model, named-entity recognition, relationship extraction knowledge fusion and knowledge graph generation. This paper takes a tight sandstone block as an example to build a knowledge graph in the field of fracturing, establish a fast and accurate inference mechanism for predicting fracturing effects, and explore intelligent optimization methods and technologies for fracturing schemes based on the knowledge graph (Fig. 1).

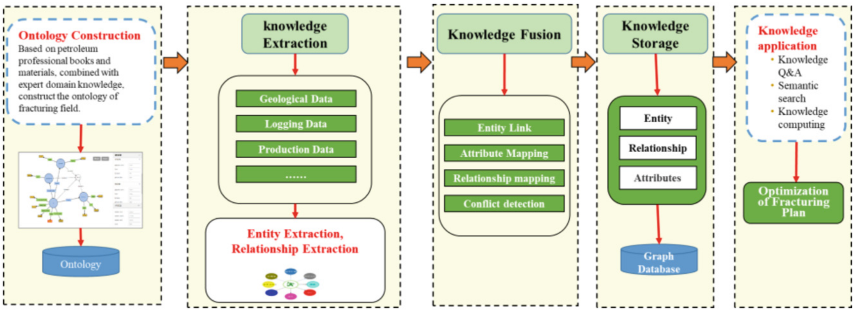


**Fig. 1.** Construction process of knowledge graph in fracturing field

## 2.1 Construction of Fracturing Domain Ontology

Ontology is a concept derived from philosophy, which refers to "a formal and detailed explanation of a shared conceptual system". Ontology construction is the basis for the construction of knowledge graph, which consists of three parts: objects, activities and features [8]. Activities act on objects, and features are used to describe activities and objects. The most basic ontology includes concepts, concept levels, attributes, attribute value types, relationships, relationship domain of a function concept sets, and relationship value domain concept sets, which can be constructed by combining top-down and bottom-up methods [3, 9, 10].

### 2.1.1 Principles of Ontology Construction

① In the lifecycle of an activity, it may involve several objects, and the activity may also generate some new objects.
② Usually, a large activity may consist of several sub activities.
③ Features are not only descriptions of objects and activities, but also descriptions of "relationships between objects", "relationships between objects and activities", and "relationships between activities".

For example: The pore type of XX reservoir is mainly intergranular dissolution pore, which describes the specific pore type of XX reservoir.

### 2.1.2 Definition and Classification of Professional Terms

① Organize concepts and unique expressions within the field of logging, add relevant information, and determine synonyms.

Consider from both horizontal and vertical directions: the horizontal direction refers to the breadth of the field, such as what concepts need to be included in the logging field. The vertical direction refers to the depth of the field, and it is necessary to consider which granularity of professional vocabulary to include. If the granularity is too small, it will lead to low efficiency and overload, and if the granularity is too small, it will cause information omission.

② Determine concepts, attributes, and relationships, and classify professional vocabulary according to professional knowledge and hierarchy.

For example: The logging work acts on the well, generating a logging curve, and the low value of the GR curve represents the reservoir. Activity: Logging, target: well, logging curve, GR curve, reservoir, characteristic: reservoir characteristic is low GR curve.

③ Integrate the previously defined word concepts and semantic relationships to form an ontology (Fig. 2).

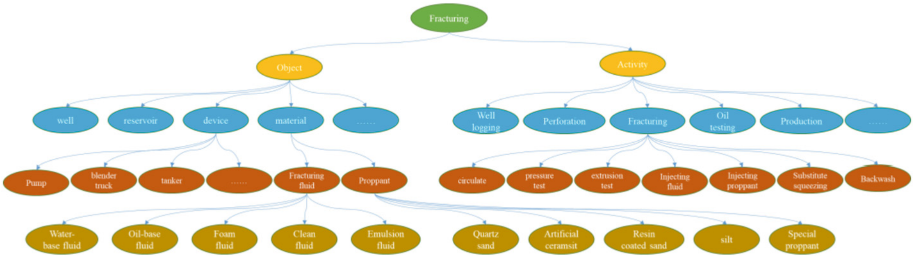


**Fig. 2.** Structural Description of Fracturing Domain Ontology

## 2.2 Construction of Knowledge Graph

### 2.2.1 Information Extraction

After the completion of knowledge ontology modeling in the fracturing field, in-depth knowledge extraction and management can be carried out on multi-source heterogeneous knowledge achievements, experience knowledge and structured data in the block, mainly including named-entity recognition, relationship extraction, attribute classification and attribute value extraction [11].

The goals of knowledge extraction include: entities of various categories, relationships between entities, and attribute values of entities. Knowledge extraction can be simply called triplet extraction.

For example: The main distribution range of porosity in the XX fracturing section of XX well is (8.0–16.0)%, with an average value of 10.7% and a median value of 11.5% (Table 1).

**Table 1.** Example of triplet extraction

| Entity | Attribute | Attribute Value |
|---|---|---|
| XXFracturing section | Porosity distribution range | (8.0–16.0)% |
| XXFracturing section | Average porosity | 10.7% |
| XXFracturing section | Median porosity | 11.5% |

### 2.2.2 Knowledge Fusion

Knowledge fusion is used to match and merge potential identical entities in the knowledge mapping results, including entity matching, attribute alignment, conflict resolution, etc. By defining matching similarity functions and thresholds for different types of entities, matching and fusion functions for the same entity are completed [12, 13]. The fusion of knowledge in the field of fracturing includes the fusion of concept layer and entity layer. The fusion of concept layer is mainly based on the extension of knowledge in the fracturing domain ontology, while the fusion of entity layer adopts entity linking technology (Fig. 3).

For example: < HuaXX Fracturing Section • Porosity • (8.0 ~ 16.0)% >
< huaXX fracturing section • porosity • (8.0–16.0)% >

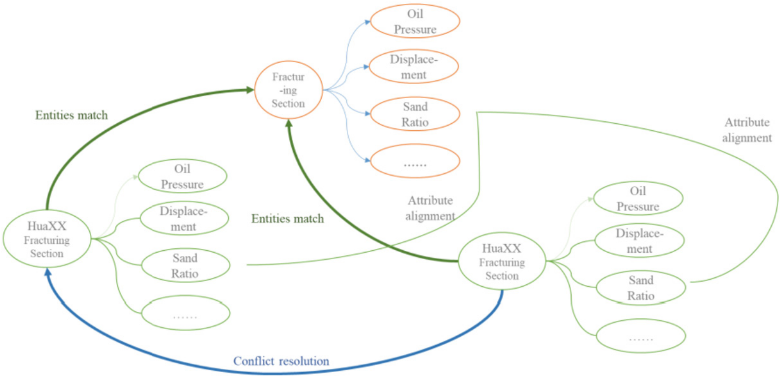


**Fig. 3.** Example of knowledge fusion

## 3 Graph Embedding Algorithm

Graph embedding is to map the original graph data (usually a sparse high-dimensional adjacency matrix) to a low latitude, dense, real value Vector space. The resulting representation vector can be used for downstream tasks, such as node classification, link prediction, visualization, etc., which can well solve the problem that graph data is difficult to input into machine learning algorithms efficiently [14]. Graph embedding generally includes three types: node embedding, edge embedding, and graph embedding.

The design of fracturing schemes has the characteristics of complex processes and multiple data analysis factors. The data types used in the model are diverse, including geological data, logging data, fracturing construction data, formation testing, and oil and gas production data and reports. The primary task of the knowledge graph is to embed the knowledge graph, embedding the knowledge base into a low dimensional space (such as 10, 20, and 50 dimensions). After obtaining the vector, various mathematical tools can be used for analysis.

Graph embedding algorithms of knowledge graph include TransE algorithm, DeepWalk algorithm, Node2vec algorithm, SDNE algorithm [14–16], among which the most classic is TransE model, which is also the algorithm used in this paper.

### 3.1 TransE Algorithm

TransE is an algorithm proposed to solve the relational data in the knowledge graph. Its function is to translate triples into graph embedded word vectors. The triple form is (head entity, relationship, tail entity), where the head entity and tail entity are collectively referred to as entities. For simplicity, h, r, and t are used to represent triples. Where h represents the head entity, r represents the relationship, and t represents the tail entity.

The core idea of the TransE model is to vectorize the relationships and entities in the knowledge graph, and adjust the vector representation of h, r, t through continuous learning to make the sum of h and r vectors equal to t as much as possible, that is, h + r = t. It is suitable for low dimensional dense representation of each knowledge in the knowledge graph.

TransE model:
Given a set of triples S, each triplet is represented as (h, r, t), satisfying h + r = t.

The distance formula is defined as:

$$d(\boldsymbol{h}, \boldsymbol{l}, \boldsymbol{t}) = \|\boldsymbol{h} + \boldsymbol{l} - \boldsymbol{t}\|_2^2 \tag{1}$$

Define the loss function:

$$\mathcal{L} = \sum_{(h,l,t)\in S} \sum_{(h',l',t')\in S'(h,l,t)} [\gamma + d(\boldsymbol{h} + \boldsymbol{l}, \boldsymbol{t}) - d(\boldsymbol{h}' + \boldsymbol{l}, \boldsymbol{t}')]_+ \tag{2}$$

Among them, [x] + indicates that the original value is taken if it is greater than 0, and 0 is taken if it is less than 0.

$$S'_{(h,l,t)} = \{(h', l, t)|h' \in E\} \cup \{(h, l, t')|t' \in E\} \tag{3}$$

The above equation represents the damaged triplet, where the head or tail entities are replaced by random entities as the control group. When training the model, the original Triplet loss function is expected to be smaller, and the damaged triplet loss function is expected to be larger.

### 3.2 DeepWalk Algorithm

DeepWalk algorithm pioneered the idea of embedding word embedding in the NLP field into the graph structure. Based on Word2vec model, each node in the graph is represented by a low dimensional vector. In order to meet the input requirements of the skip gram model, the DeepWalk algorithm utilizes the random walk sampling method to sample each node. The sampled node sequence is treated as a statement in NLP and fed into the skip gram model for training, thereby obtaining the d-dim low dimensional embedded representation of each node.

### 3.3 Node2vec Algorithm

The Node2vec algorithm is a graph embedding algorithm that utilizes deep learning algorithms to represent nodes in a graph using low dimensional vectors (d-dim). The basic idea of the Node2vec algorithm is consistent with the DeepWalk algorithm, which samples the nodes in the graph and sends the sampled nodes as corpus into the Word2vec model for learning and prediction, ultimately obtaining embedded representations of each node; The progress of Node2vec algorithm is to use the biased random walk sampling method to balance the breadth-first search (BFS) and depth-first search DFS through the selection of parameters P and Q, so as to better obtain the local characteristics and global structure characteristics of nodes (Fig. 4).

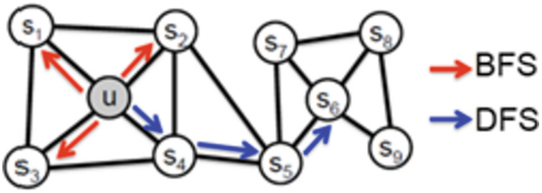


**Fig. 4.** Node2vec algorithm

## 4 Optimization Model for Fracturing Scheme

Comprehensively considering geological data, logging data, fracturing construction data, perforation data, post fracturing production and other fracturing related data, through machine learning technology, an intelligent fracturing analysis and prediction model is established to realize the prediction of fracturing effect and optimization of fracturing construction parameters, significantly improve the working efficiency and quality of fracturing design, and then improve the success rate of fracturing construction (Fig. 5).

### 4.1 Extraction of Multidimensional Feature Parameters of Curves

For logging curves and fracturing construction curves, multi-dimensional feature extraction and fusion are carried out, and deep reinforcement learning is used to characterize multi information features. The multi-dimensional features of curves are combined with relevant business data to provide a feature basis for accurate modeling. Mainly the numerical features (maximum, minimum, mean, and median), morphological features

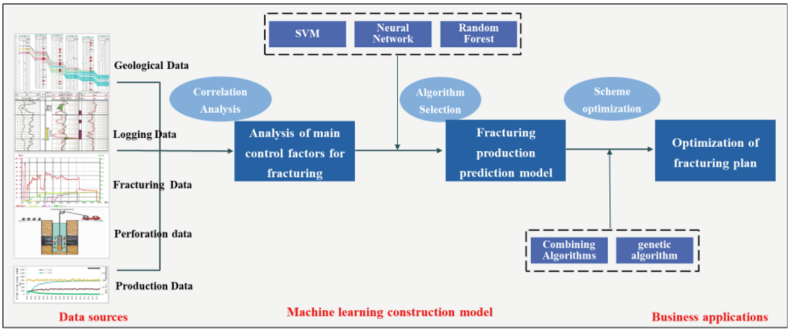


**Fig. 5.** Optimization of fracturing scheme technology roadmap

(variance, standard deviation, baseline offset, etc.), and combination relationship features between curves [8]. The selection of logging curves includes SP, GR, AC, DEN, CNL, etc., while the fracturing construction curve includes oil pressure, casing pressure, displacement, sand ratio, etc (Fig. 6).

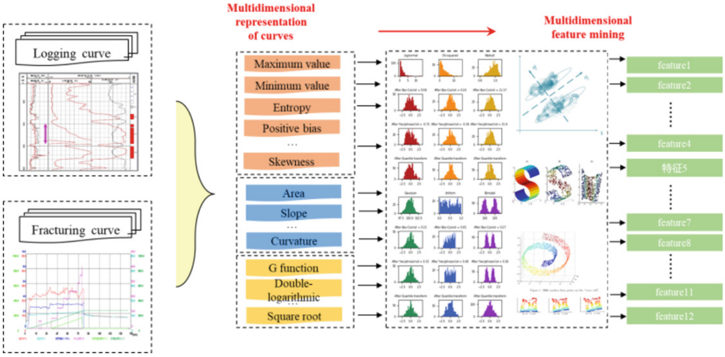


**Fig. 6.** Multidimensional feature parameter extraction of curves

## 4.2 Knowledge Feature Parameter Extraction

The model considers the multi-attribute and multi-scale features of logging data from different perspectives, and automatically extracts these features using deep learning technology. Using the knowledge of block geological characteristics, adjacent well characteristics and fracturing interval characteristics included in the established knowledge graph, the knowledge graph is organically combined with the depth neural network to establish a learning model to achieve the extraction of formation characteristics. Based on the distributed representation model of the knowledge graph, knowledge such as adjacent well information, geological stratification and production data in the map are vectorized to realize the numerical representation of knowledge (Fig. 7).

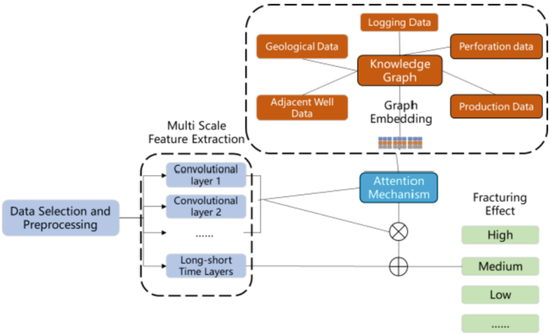


**Fig. 7.** Schematic diagram of multi-scale feature extraction for fracturing effect prediction model

### 4.3 Analysis of Main Control Factors

Using Pearson correlation coefficient matrix to eliminate the interference of artificially selected parameters, based on geological data, logging data, fracturing construction data, perforation data, post fracturing production situation and other data, the degree of influence of each parameter on production capacity is quantified through correlation coefficient, and the main influencing factors on fracturing effect are identified, providing reference for subsequent model training.

The main calculation formulas involved in Pearson's algorithm are:

$$r_{xy} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{\left[N\sum X^2 - \left(\sum X\right)^2\right]\left[N\sum Y^2 - \left(\sum Y\right)^2\right]}} \tag{4}$$

X: Reservoir thickness, permeability, oil saturation, porosity, well spacing, brittleness index, number of fracture clusters, sand addition strength, amount of injected fluid.

Y: Capacity.

**Table 2.** Parameter correlation grading table

| Correlation coefficient | Correlation |
|---|---|
| $0.8 < \|r\| \le 1$ | Extremely strong correlation |
| $0.6 < \|r\| \le 0.8$ | Strong correlation |
| $0.4 < \|r\| \le 0.6$ | Moderate correlation |
| $0.2 < \|r\| \le 0.4$ | Weak correlation |
| $0.0 < \|r\| \le 0.2$ | Extremely weakly correlated or uncorrelated |

### 4.4 Prediction of Fracturing Effect

Feature extraction is carried out based on the original logging curve and fracturing operation curve to obtain the key factors affecting the stimulation effect, and then it is integrated with the basic parameters of the relevant reservoir. The stimulation effect is classified through the full connection layer, and finally n results with excellent stimulation effect are output. Business experts can ultimately obtain the optimal fracturing construction parameter plan based on on-site experience and considering economic factors (Fig. 8).

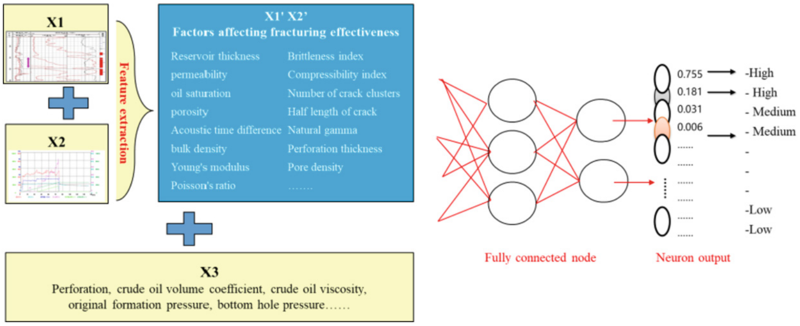


**Fig. 8.** Prediction of fracturing effect

# 5 Application Cases and Effects

In this study, more than 500 fracturing oil wells in a tight sandstone block are taken as objects to build a knowledge graph in the field of fracturing; Based on oil testing and production data and historical production, predict the fracturing stimulation effect and optimize the fracturing engineering parameters. The data used includes geological data, experimental analysis, logging, oil testing, geological stratification, and logging interpretation results (including but not limited to the parameters in Table 2). The post fracturing effects are divided into five categories: high yield, medium high yield, middle yield, medium low yield, and low yield.

## 5.1 Analysis of Main Control Factors

Based on geological data, logging data, fracturing construction data, perforation data, and post fracturing production data, the degree of influence of each parameter on production capacity is quantified through correlation coefficients to identify the main influencing factors on fracturing effectiveness, providing reference for subsequent model training. The calculation results indicate that factors such as reservoir thickness, oil saturation, permeability, number of fracture clusters, and half length of fractures have a significant impact on the fracturing effect (Fig. 9).

## 5.2 Prediction Effect of Validation Set

Using 736 fracturing sections from over 550 wells as the entire dataset, 80% of the wells were randomly selected as the training set and 20% as the validation set. The training

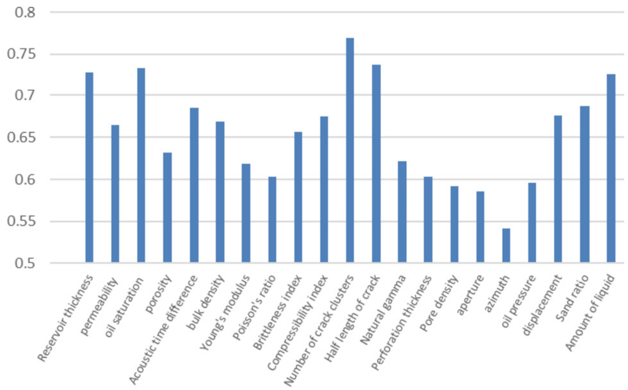


**Fig. 9.** Analysis of main control factors for fracturing effect

effect of the model is quantitatively evaluated using three indicators: precision rate, recall rate, and F1 value of artificial intelligence. The comparison with the expert explanation results shows that the model prediction accuracy of the validation set is 91.3%, the recall rate is 81.2%, and the F1 value is 85.95%. The validation effect is ideal and can meet production requirements (Table 3).

**Table 3.** Comparison and analysis table of intelligent prediction results and actual measurement results of the verification set

| Actual results | Intelligent prediction | | | | |
|---|---|---|---|---|---|
| | High | medium to high | medium | Low to medium | Low |
| High | 127 | 9 | 5 | 0 | 0 |
| medium to high | 13 | 186 | 11 | 2 | 0 |
| medium | 1 | 3 | 139 | 5 | 0 |
| Low to medium | 0 | 2 | 4 | 117 | 3 |
| Low | 0 | 0 | 1 | 5 | 103 |

### 5.3 Optimization and Verification of Fracturing Design Plan

Taking the fracturing section of xxx well (1273.6–1288.5 m) as an example, the fracturing effect prediction and optimization of fracturing construction parameters are carried out for this section, and a specific fracturing design plan is finally provided. In the optimal plan, the maximum construction pressure is 37.1MPa, the total amount of liquid entering the well is 325.6 m$^3$, the construction displacement is 2.5–3.5 m$^3$/min, and the post compaction effect is classified as high yield.

After the fracturing construction of this fracturing section, it was concluded that the total amount of fluid entering the well was 312.7m$^3$, with a construction displacement of 2.7–3.6 m$^3$/min, a maximum construction pressure of 36.5MPa, and an instantaneous ground pump stop pressure of 12.8Mpa. Finally, the sand addition was completed according to the design.

Using small-scale fracturing analysis data, the main fracturing construction was simulated, and the simulation results were basically consistent with the design plan.

Compared with the daily gas production before and within 30 days after the measures were taken, the gas production of the well significantly increased after the pressure was applied, and the production capacity was classified as high yield (Fig. 10 and Table 4).

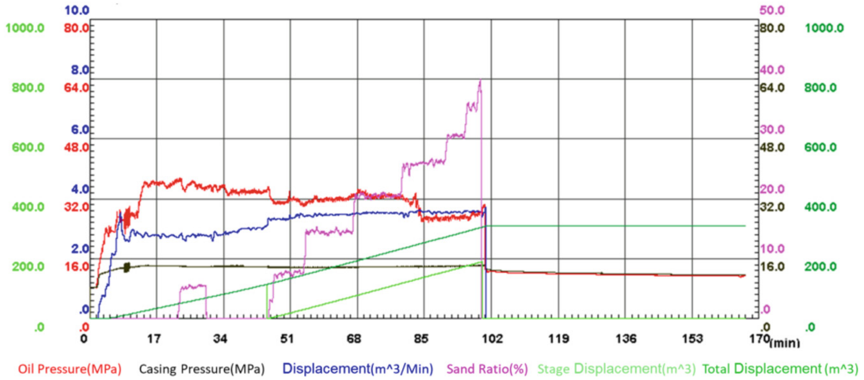


**Fig. 10.** Well xxx (1273.6–1288.5m) fracturing construction curve

**Table 4.** Comparison table between main fracturing fracture design and actual practice

| Name | Design | Actual |
|---|---|---|
| Crack length (m) | 155.7 | 146.1 |
| Support joint length (m) | 137.5 | 128.2 |
| Dynamic crack height (m) | 54.2 | 49.4 |
| Support crack height (m) | 47.8 | 43.4 |
| Crack top depth (m) | 1244.7 | 1243.5 |
| Crack bottom depth (m) | 1298.9 | 1292.9 |

## 6 Conclusion

In this paper, the knowledge graph of fracturing field is constructed by combing the knowledge system of fracturing field; extracting fine-grained feature parameters from multiple dimensions and scales, and using graph embedding technology to represent the entities, relationships, and attributes associated with these feature parameters as vector feature maps; Integrate expert knowledge with artificial intelligence to build a fracturing effect prediction model and optimize the fracturing plan. The application results show that the predicted production capacity matches the actual production capacity by over 91%, and the design work efficiency of the fracturing scheme has been increased by more than 20 times. The research results can provide scientific basis for predicting fracturing effects and optimizing fracturing engineering parameters, greatly improving the efficiency and quality of fracturing design, and improving the success rate of fracturing construction. It should be noted that although Knowledge graph can be widely used, there are still technical challenges. ① The field of fracturing has a large amount of data and exists in the form of multi-source heterogeneity, with varying data standards and quality, as well as strong professionalism. ② For different types of oil and gas reservoirs and different data bases, a series of Knowledge graph need to be established pertinently, and business experts and AI experts are closely combined to deepen research.

## References

1. Chen, Y., Liu, Z.Y., Chen, J., et al.: History and theory of mapping knowledge domains. Stud. Sci. Sci. **26**(3), 449–460 (2008)
2. Zhang, F.L., Zhang, E.L., Xiang, Y.H., et al.: Application of knowledge atlas technology in knowledge management of oil and gas exploration and development. China CIO News **2**(1), 128–131 (2020)
3. Xia, L, Wu, B.Y., Liu, L.X., et al.: Question-answering using keyword entries in the oil & gas domain. In: 2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), Shenyang (2020)
4. Li, Z., Wang, Z.F., Wei, Z.C., et al.: Cross-oilfield reservoir classification via multi-scale sensor knowledge transfer. Proc. AAAI Conf. Artif. Intell. **35**(5), 4215–4223 (2021)
5. Liang, K., Ren, Y., Shang, Y., et al.: Review on research progress of deep learning driven knowledge tracking. Comput. Eng. Appl. **57**(21), 41–58 (2021)
6. Wu, T., Qi, G., Li, C., et al.: A survey of techniques for constructing Chinese knowledge graphs and their applications. Sustainability **10**(9), 32–45 (2018)
7. Zhang, X., Jiang, L.: Review of research about ontology conception. J. China Soc. Sci. Tech. Inf. **26**(4), 527–531 (2007)
8. Liu, G., Gong, R., Shi, Y., et al.: Construction of well logging knowledge graph and intelligent identification method of hydrocarbon-bearing formation. Petrol. Explor. Dev. **49**(3), 502–511 (2022)
9. Zhu, Y.Q., Zhou, W.W., Xu, Y., et al.: Intelligent learning for knowledge graph towards geological data, pp. 1–13. Hindawi Publishing Corporation Scientific Programming (2017)
10. Ji, S., Pan, S., Cambria, E., Marttinen, P., Philip, S.Y.: A survey on knowledge graphs: representation, acquisition, and applications. IEEE Trans. Neural Netw. Learn. Syst. **33**(2), 494–514 (2021)
11. Kejriwal, M.: Domain-Specific Knowledge Graph Construction, pp. 57–58. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-12375-8
12. Wang, H, Zhang, F, Zhao, M, et al.: Multi-task feature learning for knowledge graph enhanced recommendation. In: The World Wide Web Conference, pp. 2000–2010 (2019)
13. Wang, X., Wang, D., Xu, C., et al.: Explainable reasoning over knowledge graphs for recommendation. Proc. AAAI Conf. Artif. Intell. **33**(01), 5329–5336 (2019)
14. Yuan, L., Li, X., Wang, X., et al.: Graph embedding models: a survey. J. Front. Comput. Sci. Technol. **16**(01), 59–87 (2022)

15. Cai, H., Zheng, V.W., Chang, K.C., et al.: A comprehensive survey of graph embedding: problems, techniques, and applications. IEEE Trans. Knowl. Data Eng. **30**(9), 1616–1637 (2018)
16. Goyal, P., Ferrara, E.: Graph embedding techniques, applications, and performance: a survey. Knowl.-Based Syst..-Based Syst. **151**, 78–94 (2018)

# Five Reservoir Fields 4D Visualization and Dynamic Analysis Based on WebGL and GPU Acceleration

Jian Duan[1], Yi-ran He[1](✉), Yong-bin Bi[1], Cheng-lin Yu[1], Li-li Qu[1], Rui-jie Geng[1],
Ya-hui Shi[2], Chun-jia Min[2], and Bo Xi[3]

[1] PetroChina Jidong Oilfield Company, Tangshan, China
npq_hyr@petrochina.com.cn
[2] Kunlun Digital Technology Co. Ltd, Beijing, China
[3] Geosocket (Beijing) Technology Co. Ltd, Beijing, China

**Abstract.** Based on the dream cloud platform, and on the Jidong oilfield basis of basin-level regional lake construction, some basic requirements for oilfield production analysis have been realized. However, with the development of informatization and the deepening of research work, it is necessary to develop business characteristic scenarios such as modern oil reservoir visualization and collaborative management. Using WebGL and GPU acceleration technology, an integrated collaborative online work platform is formed that integrates the "five fields" of reservoirs 4D virtual visualization and dynamic analysis results. The results show that: ①Through the standardization of model volume analysis and storage format, online unified management and visual display of 3D reservoir numerical simulation models with a large number of grids based on WebGL and GPU acceleration technology have been realized. ②Based on the simulation results of the "five fields" of four-dimensional geological field, pressure field, saturation field, seepage field and chemical field, by mining the internal data relationship of the reservoir dynamic model, the dynamic analysis and 2D-3D correlation graph could be visualized online, providing support for deep mining of reservoir five-field model data. ③Reservoir research that IFEDC-202315248 2 integrates 2D reservoir dynamic analysis, 3D model visualization, 4D model visualization and five-field quantitative analysis is an internal fluid migration based on macro development situation analysis during the entire life cycle of the reservoir In-depth analysis of the

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
J. Lin (Ed.): IFEDC 2023, SSGG, pp. 407–423, 2024.
https://doi.org/10.1007/978-981-97-0272-5_35

rules. The platform realizes scientific research collaboration among development plate research data, software, and achievements, lies a solid foundation for high-efficiency and high-quality oil and gas exploration business work, and provides strong support for the high-level application of Dream Cloud.

**Keywords:** Dream cloud · 4D Visualization · Professional software · Integration and coordination · WebGL

## 1 Introduction

With the deepening of reservoir development and the continuous accumulation of data information and results, oilfield operating areas are facing many problems in data management. In terms of exploration, development and production information management and application, due to many sources of data information, they are currently stored in data lakes, established databases, or in the hands of project team technicians. The research results are not stored and managed in a unified manner, and most of them belong to stand-alone display and application, data security cannot be guaranteed, resulting in time-consuming and laborious data retrieval in follow-up research, making it difficult to promote research progress. As an important part of reservoir research and an important means of reservoir dynamic analysis, the reservoir numerical simulation technology [1], the traditional reservoir numerical simulation research model is mostly based on the numerical simulation software commonly used in the industry, such as Eclipse, CMG, etc., use workstations to simulate models, and because these numerical simulation softwares are stand-alone applications, most of them are expensive and have limited licenses. At the same time, the software involves a wide variety of parameters and complex operations, requiring a high level of professionalism for users. The amount of data in the numerical simulation results model is huge, and the number of grids often reaches hundreds of millions. After the calculation is completed, the simulation results can only be viewed offline on a single computer. This makes it difficult to achieve online sharing of reservoir numerical simulation results and in-depth analysis based on result model volume data. The numerical simulation software used in the fine drawing work in the oil field is not unified, and is scattered and stored in the database or in the hands of technicians. The high repetition rate of data storage and the serious multi-version phenomenon lead to the inability to achieve unified management and viewing of digital simulation results, poor sharing of fine drawing results, and data security cannot be guaranteed. At the same time, because there is no research environment for in-depth exploration of the relationship between various data in the reservoir model, and a complete online information-based reservoir tracking and analysis system, the analysis methods are not standardized and unified, reducing the efficiency and accuracy of reservoir research.

Based on the above problems, this paper uses WebGL technology to establish an online model visualization window to realize the professional display of reservoir geological model and reservoir digital model in 3D space in a web browser. It can also dynamically display the changes of the pressure field, saturation field, geological field, chemical field and streamline field in the reservoir development process in the reservoir numerical simulation results. Through the GPU acceleration technology, the 3D geological model and 4D numerical simulation model of the reservoir are drawn and displayed online in real time, satisfying the rapid online mapping of massive grid model data volumes. Through the GPU acceleration technology, the 3D geological model and 4D numerical simulation model of the reservoir are drawn and displayed online in real time, satisfying the rapid online mapping of massive grid model data volumes. On this basis, the platform can also carry out in-depth data mining on the volume data of the numerical simulation model, and analyze the internal relationship between different attribute models through complex mathematical operations based on specific sub-layer combinations, and map the results in real time. The platform integrates one-dimensional data management, two-dimensional reservoir dynamic analysis, three-dimensional model visualization, four-dimensional five-field quantitative analysis, and conducts in-depth analysis of internal fluid migration laws based on macro development situation analysis during the entire life cycle of the reservoir, to achieve a more comprehensive and in-depth study of the reservoir.

## 2 Online 3D and 4D Visualization of Reservoir Model Based on WebGL Technology

Researchers in different fields at home and abroad are actively using WebGL technology to explore multi-dimensional, multi-information, and multi-scenario integration platforms.Gao Yawei [2] studied the commonly used open source Web Graphics Library (WebGL, Web Graphics Library) framework, realized the local browsing and measurement functions of 3D scenes, and built a low-cost 3D scene basic platform; Liu Zhao and Yang Jingang [3] used WebGL technology to convert oblique photographic photos taken by drones into 3D GIS models that can be browsed in a Web browser, and at the same time realized a number of key technologies such as rendering optimization and model compression. Over the years, the professional application fields of reservoir modeling and numerical simulation are all professional software systems based on the desktop version of graphics workstations/PC servers, and almost all of them use OpenGL programming technology. In the professional application field of reservoir numerical simulation based on WebGL, INT, the main technical service provider of foreign oil exploration data visualization software, will support it from 2022. However, currently limited to reservoir modeling data, the display efficiency cannot meet the data application requirements of tens of millions of modeling grids, and the online application of numerical simulation is currently in a blank period. The platform uses new WebGL and BabylonJS technologies to migrate traditional workstation-based enterprise applications to Web browsers. WebGL is the abbreviation of Web Graphics Library, which is a technology for rendering 2D and 3D graphics in browsers [4]. It is a JavaScript API based on OpenGL ES 2.0 and allows interactive 3D and 2D graphics to be rendered without plugins in any compatible web browser. WebGL is designed to run on the web and is supported by many modern browsers, including Chrome, Firefox, Safari, and Edge, among others. Because WebGL directly utilizes the graphics processing unit (GPU), it can provide excellent graphics rendering performance, which enables WebGL to create complex enterprise-level professional applications and provide high-quality visual experience [5]. In addition, through the secondary development of BabylonJS, the platform upgrades E&P professional applications on traditional workstations to web-based cloud applications. BabylonJS technology is a very active WebGL-based development framework in recent years. BabylonJS has powerful rendering capabilities, including support for PBR (Physically Based Rendering, physically based rendering) materials, ambient lighting, lightmaps, particle systems, shadows, reflections, refractions, HDR (high dynamic range), and even real-time ray tracing. BabylonJS is designed to be easy to learn and use, even developers who are not familiar with 3D programming can quickly start creating 3D scenes. Additionally, it provides a number of pre-built objects and systems that simplify scene creation and management.In addition, through the secondary development of BabylonJS, the platform upgrades the E&P professional application on the traditional workstation to a web-based cloud application. BabylonJS technology is a very active WebGL-based development framework in recent years. BabylonJS is designed to be easy to learn and use, even developers who are not familiar with 3D programming can quickly start creating 3D scenes. Additionally, it provides a number of pre-built objects and systems that simplify scene creation and management.

With the help of the above two technologies, developers can use the system graphics card to display 3D scenes and models smoothly in the browser, so as to realize the online space 3D display of complex geological models and the 4D dynamic display of multi-time step reservoir numerical simulation models. The platform extracts reservoir model grid coordinates, attribute information, and grid topology by analyzing data formats such as reservoir digital simulation software GRDECT and EGRID. Through the triangular grid processing of the grid plane, the attribute data of each grid is used to obtain the vertex color in the palette, and the BabylonJS technology is used to perform three-dimensional rendering based on the Web to realize the 3D display of the modeling attribute model (Fig. 1).

For four-dimensional visualization, the distribution of its spatial form is composed of ordered grids in the three directions of I, J, and K. The difference is that the numerical simulation results have changes in the time dimension, thus it is necessary to visualize the numerical simulation model. First, use the same spatial rendering method as the geological model to draw the model body of the first time step in the numerical simulation results, and the attribute value in it is the value of each grid on the current attribute model of the first time step. After the model volume rendering of the first time step is completed, the model mesh remains unchanged, and the properties change with time. Therefore, it automatically increments according to the time step, obtains the attribute value of the model in the next time step, completes the drawing of the model volume in the second time step, and so on, finally realizes the four-dimensional visualization function of the reservoir digital simulation results.

It should be noted that the numerical simulation process often undergoes complex grid attribute screening, and some invalid grids that do not participate in reservoir fluid migration are eliminated through the cut-off function of the numerical simulation software, so as to reduce the number of grids. For the purpose of modulo computing grid numbers, however, this process has increased the difficulty of online 4D visualization of the numerical simulation model, and the broken model meshes are distributed in a

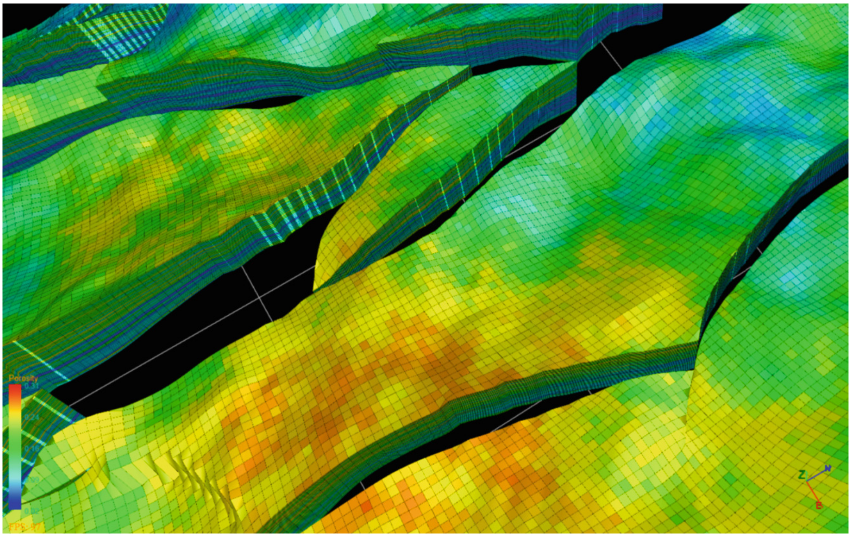


**Fig. 1.** 3D map of geological attribute model based on Web 3D rendering

messy manner. This is a challenge to the visualization ability of the online 3D window. Through WebGL and BabylonJS technology, the platform can visually display the fractured reservoir model and also have a good model rendering effect.Model volume compression storage based on data structure and algorithm optimization.

## 3 Model Volume Compression Storage Based on Data Structure and Algorithm Optimization

As we all know, the digital model volume generated by reservoir numerical simulation has a large amount of data, and the model volume is usually expressed in three directions of I, J, and K in space [6, 7], as shown in the Fig. 2 below. For some large blocks in the Jidong Oilfield, the number of fine reservoir model grids can reach hundreds of millions. For the process of reservoir numerical simulation, the same reservoir model will calculate different attributes at the same time, such as pressure and oil content. Saturation, water saturation, gas-oil ratio, etc., the amount of data carried by each attribute model is 100 million levels. At the same time, the reservoir numerical simulation model changes dynamically with time, which means that the same attribute model with step-by-step changes, the amount of data will accumulate exponentially, and eventually the amount of data carried by the result data body of the numerical simulation of the entire reservoir is very huge [8].

In order to solve the online visualization of such a large amount of digital and analog data, we adopted the following two steps to compress data storage and improve the efficiency of model visualization.

(1) Through the indexing method of the grid skeleton (Pillar) coordinates, the data storage of each grid X, Y, and Z is compressed, and the Pillar's top-bottom linear

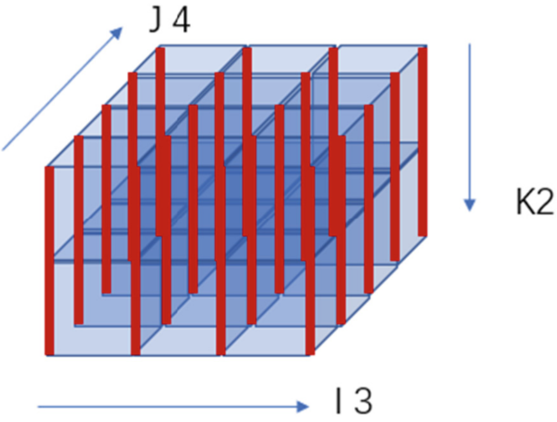


**Fig. 2.** Schematic diagram of the model mesh in I, J, and K directions

coordinate interpolation method is adopted, without affecting the visual resolution., the data storage capacity reaches more than 100 times of compression.

(2) Adopt boundary recognition “cutting” algorithm (Fig. 3). The “cutting” algorithm of boundary recognition refers to the preprocessing of removing the “hidden” internal grid in the digital-analog grid data, so as to perform 100 times lossless data compression on the model data. Due to the fast grid data “cutting” processing technology, the system can process and display reservoir numerical simulation data of hundreds of millions of grid nodes.

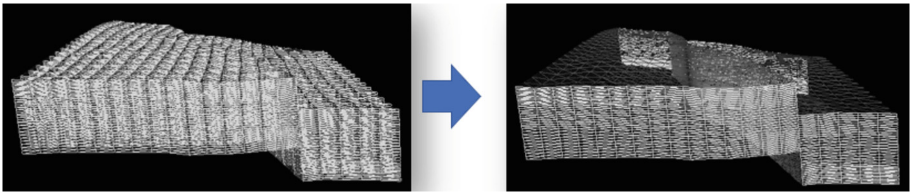


**Fig. 3.** Schematic diagram of boundary recognition “cutting” algorithm

Traditional reservoir numerical simulation operations are handled by numerical simulation software client side. The calculation speed of the model body with a large amount of data is slower, and the visualization is more difficult [9]. This platform transfers the visualization of the model from the traditional offline to online, but it will be more difficult to deal with the same volume of data online. The method of skeleton coordinate indexing can greatly compress the data storage capacity of each grid. In addition, the boundary recognition and clipping algorithm used in the model display enables online 3D space processing to realize rapid model display even for models with a large amount of data, greatly improving the efficiency of model visualization.

## 4 Based on GPU Acceleration Technology, the Efficiency of Data Calculation and Display is Greatly Improved

The five-field four-dimensional visualization and dynamic analysis of reservoirs involves a large amount of numerical simulation data calculations. The platform adopts GPU acceleration technology to improve data processing capabilities and improve the efficiency of online application of reservoir data. GPU acceleration technology is applied on the server side and the client side respectively:

(1) Parallel processing of server-side GPU

On the server side, the reservoir numerical simulation data is placed in the graphics card, and for calculation-intensive tasks, such as reservoir attribute fusion calculation (based on the reservoir numerical simulation result model body, the longitudinal mean value of a specific layer at a certain time step computing), use GPU acceleration technology to quickly calculate the mean value of each grid on the plane, and automatically fill in colors, generate contour lines, and realize online real-time mapping (Fig. 4). Through this process, researchers can determine the target according to their needs The starting and ending grid layers use a specific mean value algorithm to view the plane distribution of different attributes at different time steps of the target layer through real-time mapping results.

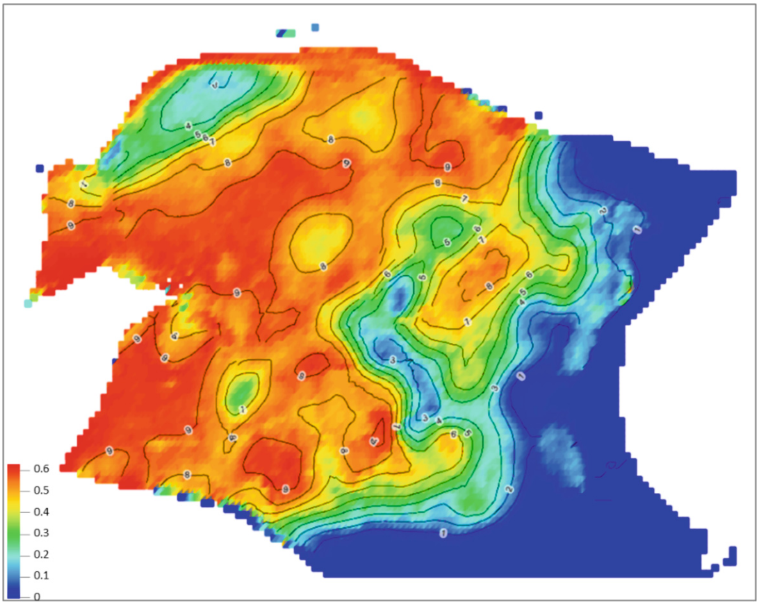


**Fig. 4.** Contour map of oil saturation plane

To realize the plane mapping of specific horizon attributes, it is necessary to set the stratification information of the sublayer, sand group and oil group first, and obtain the coordinates and values of the four corner points according to the square grid coordinates on the top surface of the sublayer.

According to the calculation of the current contour value, it is matched with the simulated attribute value of the adjacent corner point. If it falls within the value interval of the adjacent corner point, the contour control point is linearly interpolated according to the corner point coordinates and attribute value (Fig. 5). Then connect all control points and draw a plane contour map.

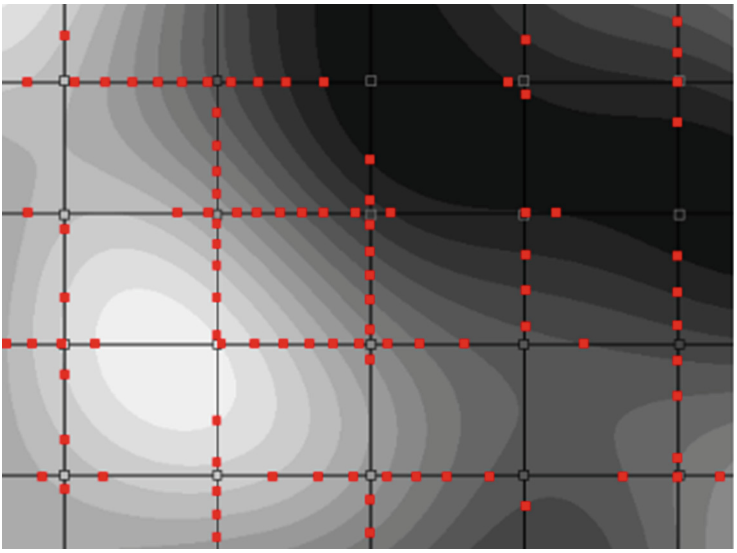


**Fig. 5.** Schematic diagram of plane contour mapping

The plane mapping of multiple grid layers requires the fusion calculation of multiple values in the vertical direction (that is, in the K direction). The mean value algorithms involved mainly include the following:

(A) Arithmetic mean: the mean obtained by simply summing and dividing by the number.

(B) Harmonic mean: Applicable to the attribute of log normal distribution, it is a kind of mean, which is the reciprocal of the arithmetic mean of the reciprocal of each sign value, also known as the reciprocal mean, simple harmonic mean calculation formula as follows:

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum \frac{1}{x}}$$

(C) Geometric mean: the square root of the multiplication of each variable value, the formula is as follows:

$$H = \sqrt[n]{a * b * c * \ldots * n}$$

(D) RMS (Root Mean Square): Also known as the square mean, it refers to the arithmetic square root of the mean of the square of a set of data. The formula is as follows:

$$Q_n = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}} = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}}$$

In addition to drawing plane distribution maps based on GPU acceleration technology, it is also possible to map the dominant seepage channels of specific sublayers in the reservoir. The dominant seepage channels are generated based on the model of porosity or permeability changing with time. The calculation method is to divide the porosity (or permeability) value of a grid at the current time node by the porosity (or permeability) value of the first day of simulation, so the dominant seepage channel plan also changes with the simulation time, based on This formula calculates the value of each grid, and finally generates contour lines and plane maps from the values of all grids on the current plane (Fig. 6), which deeply reflects the fluid migration law in the development and production process.

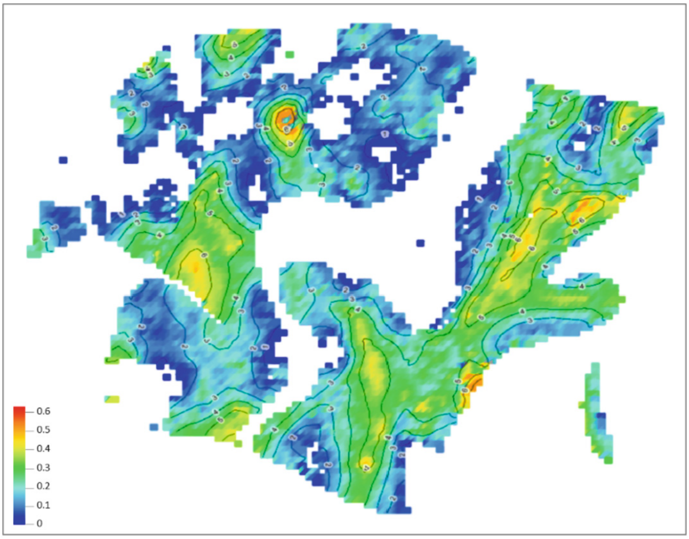


**Fig. 6.** Schematic diagram of dominant seepage channels calculated according to the porosity simulation model

HTML5 technology is the latest version of HTML, a language used to construct and present Internet content. It has become a new standard for Internet development technology [10]. With the blessing of HTNML5 technology, browsers have become a more mature platform that supports complex applications. This technology can complete operations such as drawing and rendering of two-dimensional graphics by calling simple functions, greatly enhancing the performance of Web pages. The technological innovation of the four-dimensional reservoir visualization and dynamic analysis platform is the use of HTML5 WebGL technology to realize the online application of reservoir data with tens of millions of grids. That is, how to realize the online display and data calculation of massive geological model data in the user's browser of a common commercial PC with a bandwidth of 100 megabytes. The technical solutions of this platform include the following:

(A) For the 3D visualization of the properties of the rock physics model, make full use of the shader technology of WebGL to transfer the properties of the grid to the graphics card of the GPU at one time. Model rendering (triangle mesh faces and attribute color interpolation) is done quickly using the client GPU fragment shader.
(B) For 3D grid rendering, use fragment shader boundary detection technology to realize grid boundary drawing in the client's GPU, which is more than a thousand times more efficient than conventional line segment 3D drawing methods
(C) For the drawing of two-dimensional contour lines, the ZBuffer buffer detection technology of the fragment shader of the GPU of the client is used for drawing, and the drawing efficiency of this method is also greatly improved. To sum up, this platform makes full use of the WebGL technology of Html5. The traditional CPU-based computing work is transferred to the client-side GPU-based computing. While the efficiency of visualization applications is greatly improved, the system resource occupancy rate is greatly reduced.

In terms of reservoir 4D visualization, based on HTML5 technology and GPU acceleration technology, the reservoir numerical simulation result model can be played dynamically, and any simulated attribute model including pressure, oil saturation, etc. can be displayed at any time step Check. Through the change of oil saturation in each small layer (Fig. 7), the distribution of remaining oil is analyzed to guide subsequent development and production.

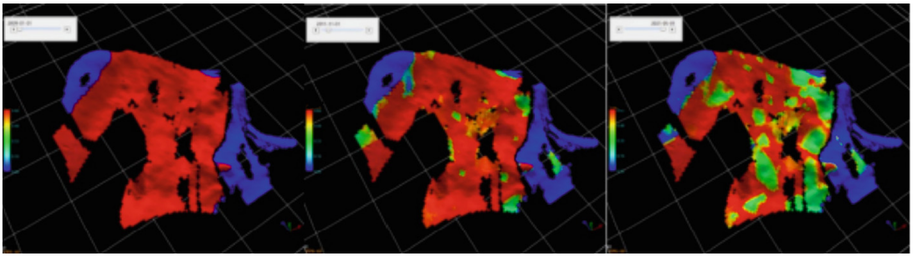


**Fig. 7.** Schematic diagram of four-dimensional visualization of oil saturation model in numerical simulation results

(2) Client GPU rendering acceleration

On the client side, for the 3D visualization and 3D visualization rendering of the reservoir model, the platform adopts GPU-accelerated graphics rendering technology and GPU-based shader (shader) technology respectively. After the vertices of the 3D model are processed by the vertex shader and the rasterized pixels are processed by the fragment shader, the reservoir properties can be quickly displayed (Fig. 8).

The reservoir geological model body is composed of millions of grids, which are arranged along the three directions of I, J, and K, and each grid is endowed with different porosity, permeability or oil saturation. The value of the attribute. Each grid is spatially a cube with eight points and six faces and irregular geometry. Since each mesh has six faces and eight vertices, the X, Y, and Z coordinates of each vertex are known, and each face has four points, and the middle diagonal points are connected and separated into two triangles.

Use GPU acceleration technology to draw a triangle surface according to the XYZ coordinates of each triangle vertex. The color of each point is determined according to the current attribute value and palette setting, and can be calculated by uniform method or interpolation method:

(A) The uniform method means that the vertex color of the triangle uniformly adopts the color corresponding to the grid attribute value.

(B) (B) The interpolation method means that the vertex color of the triangle is determined by the grid properties of the point (I, J, K). The triangle grid is drawn and submitted to the graphics card GPU. Based on the GPU acceleration algorithm, the color of the middle point is quickly calculated by GPU interpolation sure.

Using the above two methods to draw all the triangular meshes in the three directions of model I, J and K, the drawing of the geological model in three-dimensional space can be completed.

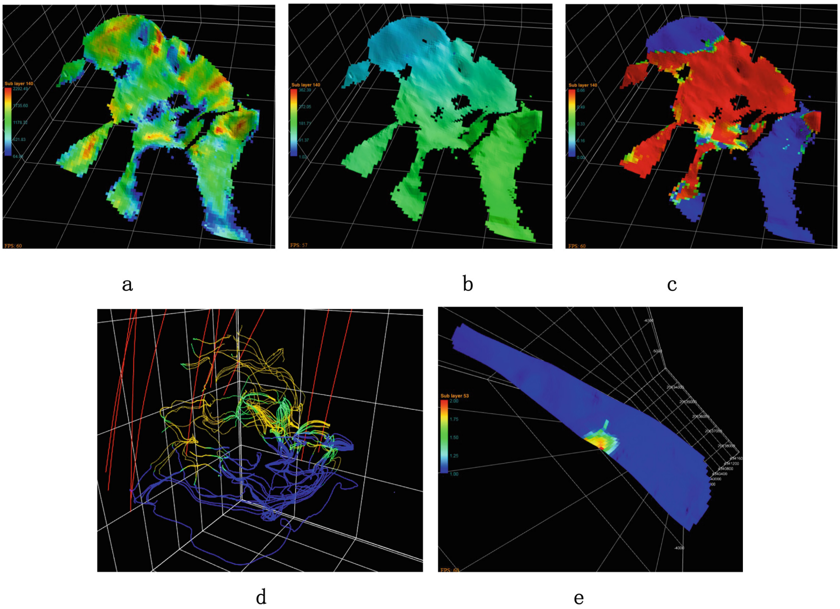


**Fig. 8.** Schematic diagram of five reservoir field (a reservoir geological field; b pressure field; c oil saturation field; d seepage field; e chemical field)

# 5 Scenario Customization Based on Model Analysis Data Body

Through the classification, analysis and conversion of multi-source data, a unified data structure is integrated, and one-dimensional to four-dimensional scene organization and management are realized based on the virtual palace platform (Fig. 9). Establish two functional modules including screening, hiding, interaction and dynamic coupling, reservoir visualization and reservoir dynamic analysis. Reservoir visualization includes wellbore visualization, seismic and geological model visualization, and reservoir virtual visualization. By analyzing the data body format of the Eclipse, CMG, tNavigator and Petrel work area model, research the modeling of the work area and the model mesh of the digital model, analyze and compile it, realize the reading, conversion and four-dimensional visual display of geological model data volume and numerical simulation result data volume, create a customized four-dimensional geological field, pressure field, saturation field, seepage field, and chemical field based on the model analysis data volume and five field quantitative visualization scenarios to carry out information-based and efficient dynamic analysis. Reservoir dynamic analysis under the new model includes development status analysis, development situation analysis, development effect evaluation, water flooding status analysis, remaining oil potential analysis and development technical policy, etc.

Through the development of professional software interfaces and data query and screening functions, the operation calling methods such as screening, display, hiding, query, dynamic and interaction are realized, which basically meets the needs of technical personnel for dynamic analysis of reservoirs and improves the efficiency of research work.

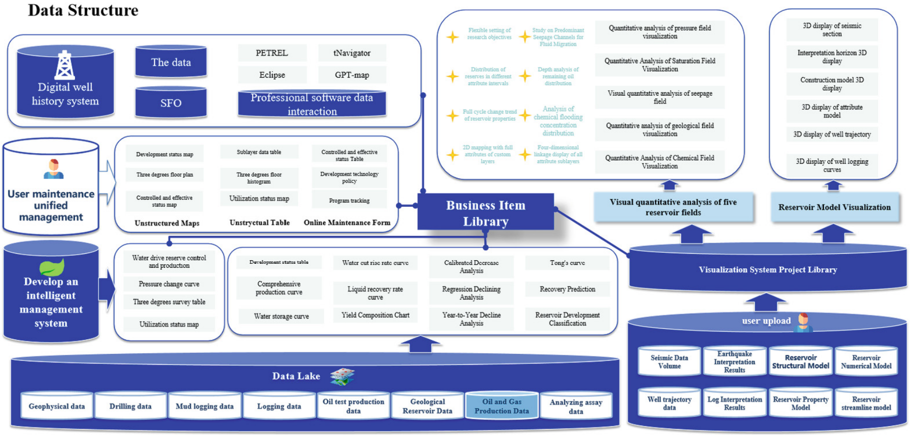

Data Structure
Digital well history system
The data
SFO
PETREL
tNavigator
Eclipse
GPT-map
Professional software data interaction
User maintenance unified management
Development status map
Three degrees floor plan
Controlled and effective status map
Unstructured Maps
Sublayer data table
Three degrees floor histogram
Utilization status map
Unstryctual Table
Controlled and effective status Table
Development technology policy
Program tracking
Online Maintenance Form
Business Item Library
Flexible setting of research objectives
Study on Predominant Seepage Channels for Fluid Migration
Distribution of reserves in different attribute intervals
Depth analysis of remaining oil distribution
Full cycle change trend of reservoir properties
Analysis of chemical flooding concentration distribution
2D mapping with full attributes of custom layers
Four-dimensional linkage display of all attribute sublayers
Quantitative analysis of pressure field visualization
Quantitative Analysis of Saturation Field Visualization
Visual quantitative analysis of seepage field
Quantitative analysis of geological field visualization
Quantitative Analysis of Chemical Field Visualization
3D display of seismic section
Interpretation horizon 3D display
Construction model 3D display
3D display of attribute model
3D display of well trajectory
3D display of well logging curves
Visual quantitative analysis of five reservoir fields
Reservoir Model Visualization
Develop an intelligent management system
Water drive reserve control and production
Pressure change curve
Three degrees survey table
Utilization status map
Development status table
Comprehensive production curve
Water storage curve
Water cut rise rate curve
Liquid recovery rate curve
Yield Composition Chart
Calibrated Decrease Analysis
Regression Declining Analysis
Year-to-Year Decline Analysis
Tong's curve
Recovery Prediction
Reservoir Development Classification
Visualization System Project Library
user upload
Seismic Data Volume
Earthquake Interpretation Results
Reservoir Structural Model
Reservoir Numerical Model
Well trajectory data
Log Interpretation Results
Reservoir Property Model
Reservoir streamline model
Data Lake
Geophysical data
Drilling data
Mud logging data
Logging data
Oil test production data
Geological Reservoir Data
Oil and Gas Production Data
Analyzing assay data

**Fig. 9.** Schematic diagram of the data architecture of the four-dimensional virtual underground palace platform

**Screening and Hiding Functions:** The geological model of the reservoir visualization module supports the functions of displaying and hiding coordinate axes, and supports the unified viewing function of seismic section data, well trajectory data and geological model data in the same window in the same 3D window. Check the seismic section, well trajectory and geological model at the same time to display it.

**Interactive Analysis Features:** The three-level survey map of the reservoir dynamic analysis module mainly conducts a comprehensive comparison and analysis of the data of the control degree of injection-production well pattern, the degree of production of water drive reserves, and the control degree of water drive reserves input from the target block after years of geological surveys. And display, to help users understand the reserve production and control state changes of the block over the years, deeply analyze the rationality of the current injection-production well pattern in the target block, improve the control degree of the oil field, and increase the reservoir production rate.

The geological model visualization of the reservoir visualization module supports the adjustment of the color of the layer and the transparency of the grid, the translation and rotation of the attribute model in the three-dimensional space, and the plane zoom-in and zoom-out and vertical stretching and compression. Adjustment of mesh transparency is supported by mouse clicks. The sliding of the mouse wheel supports zooming in and out of the model, and the moving of the mouse supports the movement of the model. And use the rapid 3D interactive interception technology to intercept local sub-regions from the overall model of the oil and gas reservoir, or draw a line segment through the model to cut the model into two parts, and the model sections of each part are displayed separately. In this way, the user can observe the key development area in the work area that he is concerned about - the attribute of the sweet area (small model of the sub-area) in the longitudinal direction.

**Dynamic Coupling Analysis Capabilities:** The five-field quantitative analysis of the reservoir visualization module uses the time step output from the reservoir numerical simulation results as the abscissa through the specific attribute model body in the numerical simulation results according to statistical analysis and other algorithms. Using the arithmetic mean algorithm, the average value of the attribute value on each grid on the specific attribute model under a single time step is calculated, use the calculated value as the ordinate to generate the average change curve of the current property in the whole area, make the 3D reservoir digital model have time attribute, and draw it into 2D plane graph, columnar statistical analysis graph, curve graph, etc. for attribute plane distribution analysis, reserve analysis and pressure strength analysis, etc., endowing it with dynamic functions, understand the overall change trend of attributes such as reservoir pressure, oil saturation, gas saturation, and water saturation during oilfield development, and finally realize the comprehensive research function of oil and gas reservoirs integrating 4D model visualization and quantitative analysis of attribute field data.

Based on the five visualization scenarios, comprehensively and accurately quantitatively analyze the internal dynamic changes of the reservoir, and can deeply mine the internal data relationship of the attribute model.

Perform simulation operations on different attributes, such as pressure, oil saturation, water saturation, and gas saturation. After the operation, you can view the dynamic change process of each attribute model. Also capable of performing simulation calculations is the reserve model. The numerical simulation software calculates the reservoir reserve model based on basic parameters such as grid volume and porosity, and combines the reserve model with other attribute models to conduct in-depth analysis of the data. Discover the changing relationship between reservoir reserves and different attributes. Taking the pressure field as an example, a certain time step is selected, and based on the current reservoir pressure model, the pressure coefficient of each grid is calculated for a specific sub-layer or oil group. The pressure coefficient of each grid is equal to the pressure value of the current time step of this grid divided by the initial pressure value of the reservoir, and the pressure coefficient of each grid will be obtained. Among them, the division of small layers, sand groups and oil groups is mainly based on the vertical grid layer of the model, that is, the grid layer in the K direction. By classifying the grid layers in the K direction, set the start and end grid layer numbers contained in each small layer, and the sand group contains multiple small layers, and the oil group contains multiple sand groups. By analogy, the mutual inclusion relationship of all levels is determined, and then the values are counted and screened, and different value ranges are divided, as the abscissa, then carry out reserve statistics on all the grids in different value range intervals, as the ordinate, the histograms and change curves of the reserves distribution in different sub-layers and different pressure coefficient intervals at different time steps are finally obtained, and a comprehensive analysis scene integrating two-dimensional, three-dimensional and four-dimensional is constructed. The same applies to the creation of other attribute scenarios (Fig. 10).

## 6 Multi-dimensional Visualization and Dynamic Analysis of Reservoir Model

Realistically displaying the multi-dimensional shape of the subsurface model through the visualization window is an important part of reservoir research [11]. The 3D visualization window established based on WebGL described above and the rapid online display of the reservoir model realized by GPU acceleration technology can be used for reservoir research. Multi-dimensional visualization of geological models can be viewed online. For example, the porosity, permeability, oil saturation and other models established in the Petrel software are superimposed on the well trajectory and logging curve at the same time to reproduce the geological conditions of the work area [12, 13]. Reservoir multi-dimensional visualization and dynamic analysis is a comprehensive use of a series of scattered and unevenly distributed data such as production dynamic data, seismic, logging and fine reservoir description research results, combined with two-dimensional, three-dimensional and four-dimensional visualization technology, through data integration, processing and comprehensive management, build an integrated collaborative operation mode for reservoir research that integrates one-dimensional data management, two-dimensional reservoir dynamic analysis, three-dimensional model visualization, four-dimensional model and five-field quantitative analysis, complete the business process combing function transplantation and function development work, establish the reservoir structure model and the spatial distribution of reservoir parameters such as permeability and oil saturation in the multi-dimensional space, basically meet the needs of reservoir research and analysis, realize digital intelligence, improve business and decision-making efficiency, save time and reduce costs. Establishing a multi-dimensional

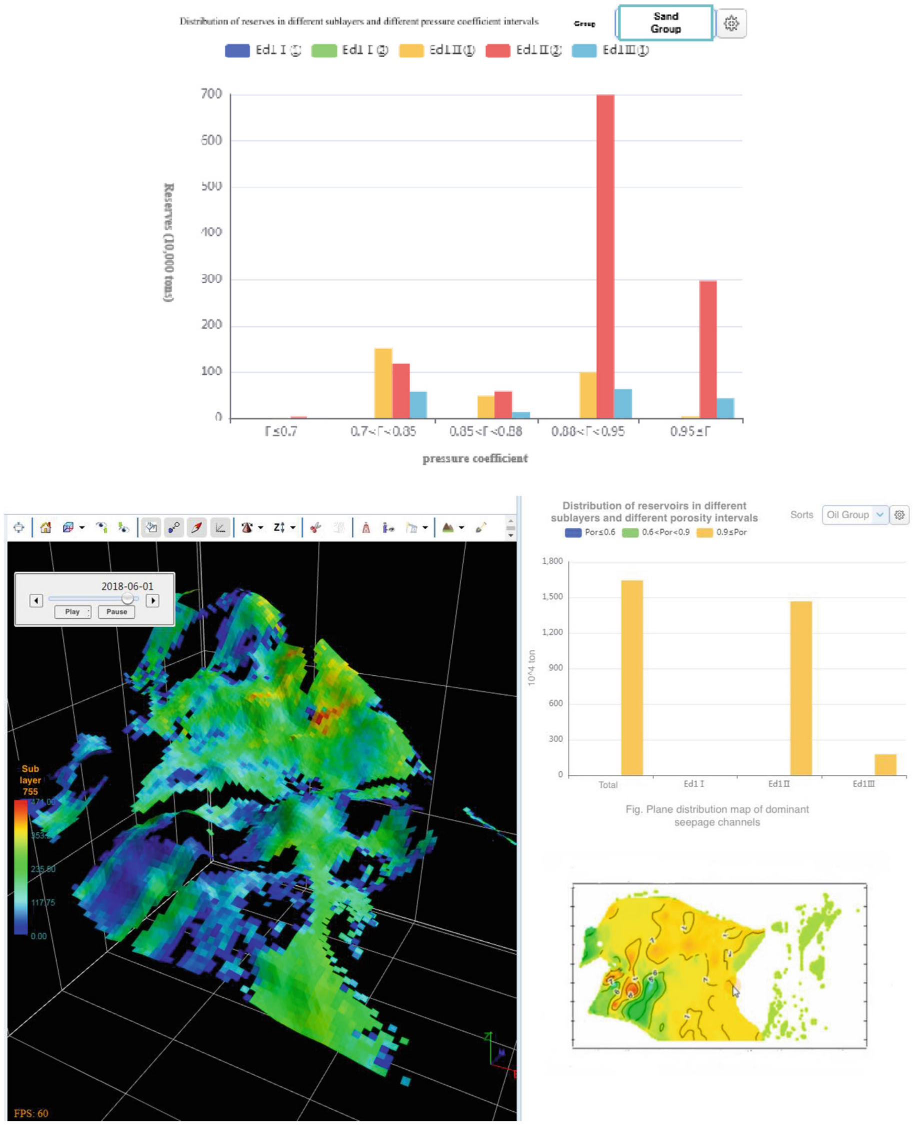


**Fig. 10.** Histogram of reserves distribution in different sub-layers and different pressure coefficient intervals

model of the reservoir not only allows the observer to feel the authenticity and fidelity of the underground space of the reservoir, but also enables spatial dynamic analysis of it (Fig. 11). For example, the geological field includes two attribute models of porosity and permeability. The pressure field is the reservoir pressure attribute model, the saturation field includes three attribute models of oil saturation, gas saturation, and water saturation, and the chemical field representation shows the simulation model of the chemical tracer concentration in the chemical flooding reservoir, the streamline shows the direction and flow rate of the fluid from the water injection well to the oil production well, which accurately reflects the distribution of injection and production between the water injection well and the oil production well. The denser the streamline, the stronger the flow field. The simulation and observation of the system can help evaluate the water injection efficiency, find the dominant flow field and channel, and guide the subsequent distribution of injection and production [14]. The analysis and research unit is refined to a single sand layer and layer, and the data and graphics are synchronized. More in-depth data analysis can be performed on each attribute model body, and the internal laws of the reservoir can be reflected through deep mining of the relationship between each model body. Observe the reservoir model from static production dynamic data to any angle, breaking the limitation of observing the reservoir model by traditional slicing method. It enables geologists and reservoir personnel to interpret and analyze exploration results more accurately, intuitively understand and study complex reservoirs, reasonably determine the spatial distribution of oil layers, and provide effective decision-making basis for tapping the remaining potential.

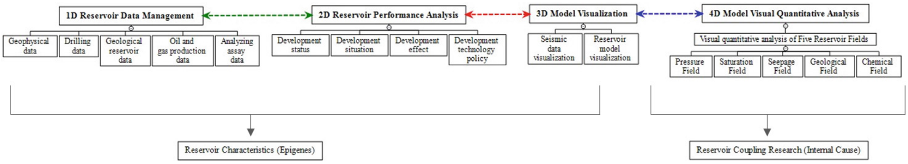

1D Reservoir Data Management
Geophysical data
Drilling data
Geological reservoir data
Oil and gas production data
Analyzing assay data
2D Reservoir Performance Analysis
Development status
Development situation
Development effect
Development technology policy
3D Model Visualization
Seismic data visualization
Reservoir model visualization
4D Model Visual Quantitative Analysis
Visual quantitative analysis of Five Reservoir Fields
Pressure Field
Saturation Field
Seepage Field
Geological Field
Chemical Field
Reservoir Characteristics (Epigenes)
Reservoir Coupling Research (Internal Cause)

**Fig. 11.** Structural diagram of the virtual underground palace project

# 7 Conclusion

(1) Under the background of digitalization, informatization, and intelligent development, the sharing of data and results has become the mainstream trend, and the results of traditional reservoir numerical simulation are mostly stand-alone applications. Not only is the amount of data large and the operation is difficult, but it is also difficult to achieve online results sharing. Based on technologies such as HTML5, WebGL, and GPU acceleration, this article intends to explore a new model for collaborative sharing of reservoir digital modeling results, and to help oilfield information construction.

(2) Reservoir numerical simulation is an important means of reservoir research, and the process of simulating reservoir dynamic changes has abundant research information. Under the surface of simple attribute changes. Using the powerful processing capabilities of computers and the various attribute model results that can be simulated by current digital simulation software [15]. Through the screening, correlation, statistics, and comparison of massive data, the internal laws of reservoir changes can be deeply analyzed to provide guidance for remaining oil research and future development plan deployment.

(3) By combining the change curve of reservoir attributes, the histogram of the relationship between reservoir attributes and reserves, the two-dimensional plane map of the reservoir, the three-dimensional geological model of the reservoir and the dynamic model of the four-dimensional numerical simulation of the reservoir, while performing online visualization of the numerical simulation results of the five fields of the reservoir. Supplemented by in-depth analysis of model volume data, a new visualization and dynamic analysis research model integrating four-dimensional reservoir visualization and five-field quantitative analysis is formed.

## References

1. Liu, P., Jiang, H.Y.: Current status and development trends of reservoir numerical simulation technology. Petrochem. Technol. **25**(03), 163 (2018)
2. Gao, Y.: Construction method and implementation of 3D scene basic platform based on open source engine. Stand. Surv. Mapp. **39**(1), 22–26 (2023)
3. Liu, Z., Yang, J.G.: Research and development of web-based 3D geographic information model management system. Railway Comput. Appl. **32**(5), 45–52 (2023)
4. Yong, C.T.: Design and research of a 3D visualization system for data centers based on WebGL. China's Manage. Inf. **26**(1), 153–156 (2023)
5. Huang, R.S., Li, C.R., Feng, L., et al.: Research on WebGL vector data 3D rendering technology based on geometry. Remote Sens. Technol. Appl. **3**, 463–468 (2014)
6. Cao, K., Pan, M., Sun, P., et al.: A method for generating 3D mesh models for reservoir numerical simulation. J. Northeastern Univ. **38**(9), 1341–1346 (2017)
7. Qi, D.S., Pei, B.L.: Theory and method of grid coarsening for reservoir models. Xinjiang Petrol. Geol. **29**(1), 91–93 (2008)
8. Liu, S.S., Jiang, H.W., Zhao, Y.P., et al.: Efficient visualization technology of geological engineering integrated reservoir model. J. Xi'an Petrol. Univ. **36**(4), 58–67 (2021)
9. Yang, H.W., Lv, D.L., Gao, S.T.: The influence of grid scale on the precision of reservoir model. J. Oil Gas Technol. **32**(3), 325–329 (2010)
10. Li, L.P.: Application of information visualization in scientific and technological monitoring. Inf. Technol. **35**(01), 150–152 (2011)
11. Guo, Z.X., Chen, P., Zhou, K.J.: Application of 3D visualization technology in oil drilling. Nat. Gas Ind. **24**(1), 60–65 (2004)
12. Ye, X.M.: Extraction of model geological information and its conversion between different grid systems. J. Xi'an Petrol. Univ. **33**(4), 44–48 (2018)
13. Li, G.T.: Research and application of key technologies for 3D visualization of reservoir geological data. China University of Petroleum (East China), Dong Ying (2017)
14. Xie, W.W.: Study on water drive development characteristics of anisotropic reservoir based on streamline method [Doctoral Dissertation]. China University of Geosciences (Beijing), Beijing (2018)
15. Wei, S., Li, L.Y., Liu, F., et al.: Research status and development trends of fine reservoir numerical simulation. Chin. Petrol. Chem. Stand. Qual. **41**(24), 78–79 (2021)

# Research on Determination Method of Oil Viscosity Based on Component Data and Machine Learning Algorithm

Yang Yu(✉), Yun-bo Li, Hao Sun, Qiang Luo, Zhao-peng Yang, Xiao-yan Geng, Zhang-cong Liu, and Xue-qi Liu

Research Institute of Petroleum Exploration and Development, CNPC, Beijing, China
yuyang2022@petrochina.com.cn

**Abstract.** Under certain conditions, when crude oil is moved by external forces, the property of internal friction generated between crude oil molecules is called crude oil viscosity. The viscosity of crude oil reflects its complex seepage state in porous media. Underground crude oil with high viscosity, always has great flow resistance in porous media, thus the flowing becomes more difficult. Oil viscosity is an indispensable key parameter in the process of dynamic analysis, reservoir engineering calculation and reservoir numerical simulation, which has critical influence on the field of well production or crude oil storage and transportation. Due to different oil viscosity, recovery approach of oil reservoirs, technical measures for storage and transportation, and the quality of oil products will be affected. The composition of crude oil is complicated, but it is mainly composed of carbon and hydrogen elements. The composition has a crucial effect on oil viscosity. Therefore, according to composition data of the actual oil sample, the determination dataset of oil viscosity is constructed together with other key parameters that affect the viscosity of crude oil within the reservoirs. Based on various machine learning algorithms, like extremely randomized trees and XGBoost, determination methods of oil viscosity based on component data and machine learning algorithms are established. In the construction process of computational model of oil viscosity, whole dataset is parted to the training dataset and the testing dataset in the ratio of 8:2. The training dataset is mainly used to determine the best hyper-parameter combination of machine learning algorithm, while the testing

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 424–433, 2024.
https://doi.org/10.1007/978-981-97-0272-5_36

dataset is used to determine the accuracy and adaptability of the corresponding method. Compared with methods such as experimental method and empirical formula method, the determination method of oil viscosity based on component data and machine learning algorithm does not require extra experimental costs and has a considerable degree of accuracy. Once the relevant input parameters are determined, the viscosity determination of multiple groups of oil samples could be completed quickly and accurately.

**Keywords:** Component Data · Machine Learning · Oil Viscosity · Determination Method

## 1 Introduction

Under certain conditions, when crude oil is moved by external forces, the property of internal friction generated between crude oil molecules is called crude oil viscosity [1]. The viscosity of crude oil reflects its complex seepage state in porous media. Underground crude oil with high viscosity, always has great flow resistance in porous media, thus the flowing becomes more difficult. During the progress of oil recovery, the viscosity of crude oil determines its seepage capacity in the formation, and during process of storage or transportation, oil viscosity also affects its flow capacity in the pipeline [2]. Thus, the accurate determination of crude oil viscosity is of great importance to productivity calculation, process design, reservoir engineering research, development planning and other works [3]. Accordingly, finding an appropriate method which could rapidly and precisely determine the viscosity of crude oil is critical.

There are many ways to determine oil viscosity, which could be generally parted to three categories: experimental approach [4], empirical formula method [5] and artificial intelligence approach [6]. As the name implies, the experimental approach involves using a combination of instruments to devise a way to actually measure viscosity value. For example, fluid viscosity could be determined using the rotating viscometer, which uses the resistance experienced by the rotor of the rotating viscometer to determine the viscosity of the liquid. However, experimental methods require us to have experimental equipment, which is usually time-consuming and costly compared to other methods. Empirical formula method is a simple method which also used to calculate oil viscosity. It usually regresses the relationship between oil viscosity and some related factors through some mathematical relation expressions. Compared with the experimental method, this method is simpler and saves time and effort. However, the calculation results are not accurate enough due to insufficient consideration factors, or the large systematic errors caused by the selection of some mathematical expressions.

In recent years, machine learning methods in the field of artificial intelligence have been extensively employed in the various fields, which can be used to solve many problems including prediction and optimization. Compared with conventional methods, machine learning algorithms are more efficient and have higher performance, and can handle many complex tasks. Therefore, this paper focuses on the determination of crude oil viscosity by machine learning method. In this paper, we first introduce the basic principles of various machine learning algorithms; secondly, the workflow of oil viscosity determination by combining component data and machine learning algorithm is introduced; next, the prediction accuracy of different machine learning algorithms is compared; finally, there is the concluding section.

# 2 Methodology

## 2.1 Decision Trees

Decision Trees (DT) is a kind of supervised machine learning algorithm, which divides the data layer by layer until all the features are divided. This process is similar to the process of leaf growth [7]. DT consists of multiple nodes which denote the properties, branches which mean the regulations and leaves which signify the results. The DT model adopts the top-down process in processing dataset. In the case of a given dataset, they will try to group and label the similarities between the data, and find the best rules for classification and regression analysis of different labels corresponding to them until the maximum accuracy is achieved.

DT are generally divided into two main patterns: classification trees or regression trees. The decision variable of classification tree is discrete, while the target variable of regression tree is continuous value. The regression tree is used in this study. In the regression problem, CART algorithm measures the split result by least square deviation (LSD), and selects the branch of the result that minimizes all possible options for splitting. The basic schematic diagram is shown in Fig. 1. The root node is the beginning of a tree, and the internal node referring splits into further nodes. The leaf node is a node that no longer splits, while the branch is the link between nodes. Whole nodes within regression tree must regressed. Through certain feature deciding criteria, we can get the final regression result.

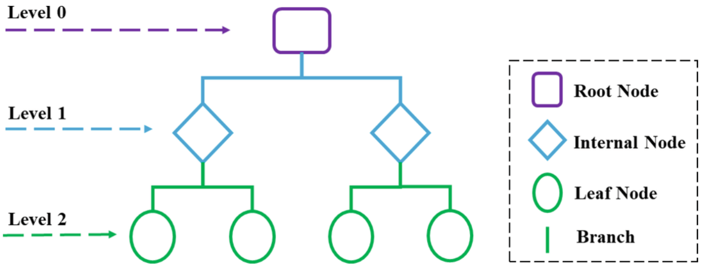


**Fig. 1.** The basic schematic diagram of DT

## 2.2 Random Forest

Random Forest (RF) is an ensemble learning (bagging) algorithm (one of parallel integrated learning approach) [8]. Its base learner is fixed as a decision tree, and multiple trees constitute a forest. Random lies in the randomness of the selection of partition attributes. Sample perturbation is also added in the way of sampling with retractions, and an attribute perturbation is introduced. That is, during training procedure of base decision tree, when selecting partition attributes, the RF first chooses one subgroup containing $N$ features at random root in candidate attribute dataset, and then selects the optimal partition attribute from this subset.

The main steps of using random forest regression algorithm for research are as follows: firstly, generate training sample $A$ and testing sample $B$, and conduct bootstrap resamples on the training sample $A$ (i.e. repeated sampling with put back), then generate $U$ new sample datasets $\{S_u, u = 1, 2, 3, \ldots\ldots, U\}$.

The second step is that each new sample set grows into a single decision tree, and $m$ ($m < M$) features randomly selected from $M$ features are used as the split feature set of the current node at the node of each tree. According to the principle of minimum impurity, the feature that meets the requirements is found in $m$ features of each node to split, and the action is repeated until the decision tree can no longer split, at the same time the $m$ is maintained constant throughout the process.

Among them, the impurity is calculated in the form of residual sum of squares, and the specific calculation formula is as follows:

$$deviance(N) = \sum_{i=1}^{N} (z_i - \overline{z})^2 \tag{1}$$

where $z_i$ is the value of class $i$ at a node, while the $\overline{z}$ is the average value.

Next, for a new observation data, the random forest uses the average value to determine the forecasting value, that is, the prediction of the observation value after the average of the prediction results of all trees. It can be expressed by the following formula:

$$F(x) = \frac{1}{q} \sum_{i=1}^{q} f_i(x) \tag{2}$$

where $F(x)$ denotes the regression results; the $f_i(x)$ represents the forecasting value of one decision tree; the $q$ means the total number of decision trees.

### 2.3 Extremely Randomized Trees

The traditional decision tree split nodes only consider the randomness of attributes and samples based on the random selection of attributes, but do not consider the impact of feature attributes on samples, thus increasing the overall bias of the algorithm and reducing the overall variance. By building a random tree, the optimal attribute is not randomly selected when the attribute is divided, but a random threshold is generated for each attribute, and then the extreme random tree-based classifier is integrated into the limit random forest algorithm. Compared with the RF algorithm, the purpose of this algorithm is to achieve better deviation and variance, so as to optimize the overall generalization error of the algorithm [9].

The key process of ET can be described as follows:

First, give the original data set $E$, the number of samples $N$, and the number of features $M$. It is worth noting that each base classifier is trained using all the samples in the classification model of the extremely randomized trees.

Next, generate a base classifier based on the CART decision tree algorithm. In order to enhance randomness, $M$ features are randomly selected from $m$ features when each node splits, and the optimal attributes are selected for each node to carry out node splitting without pruning. Perform this procedure iteratively on the split data subset until a decision tree is generated.

The step 3 is that repeat aforementioned 2 steps for $J$ iterations to generate $J$ decision trees and extremely randomized trees.

Eventually, testing samples are employed to generate prediction results from the generated extremely randomized trees, and the forecasting results of all base classifiers are counted, thus the final classification results are generated by the voting decision approach.

### 2.4 Support Vector Regression

On the basis of minimization of structural risk, support vector regression (SVR) machine can effectively improve the generalization performance of learning machine by constructing the best classification hyper-plane, and it usually has excellent performance in solving problems with small samples or nonlinear situations [10]. So, it has been widely used in many real research domains.

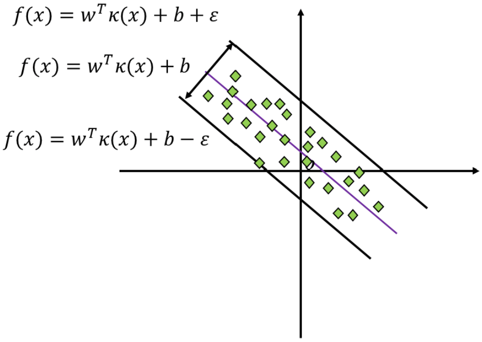

$f(x) = w^T\kappa(x) + b + \varepsilon$
$f(x) = w^T\kappa(x) + b$
$f(x) = w^T\kappa(x) + b - \varepsilon$

**Fig. 2.** Sketch Map of SVR using support vector separation

SVR could be used to classify the problems without normal linearity, and map the data to the high dimensional space by using the suitable kernel function of SVR, so that realize linear classification in the high dimensional space (as depicted in Fig. 2). The quality of kernel function mapping is related to the accuracy of final forecasting results. In the application process of SVR algorithm, the design and selection of kernel function is very important. The common types of kernel function mainly include linear kernel function, polynomial kernel function, radial basis kernel function (RBF), Laplacian kernel function and Sigmoid kernel function.

Among aforementioned kernel function, RBF is employed in following studies, as which could be expressed as Eq. (3):

$$K(x, x_1) = \exp(-\gamma \|x, x_1\|) \tag{3}$$

In above equation, $\gamma$ refers the width parameter of radial basis kernel function. Then the final SVR model is:

$$f(x) = \sum_{i=1}^{N} \left(\alpha - \alpha_i^*\right) \exp(-\gamma \|x, x_1\|) + b \tag{4}$$

### 2.5 Gradient Boosting Regression

Boosting is a kind of basic classifier (weak classifier) generation mode of ensemble learning, whose core idea is to generate a series of learners through iteration, giving high weight to those with low error rate and low weight to those with high error rate, and combining weak learners with corresponding weights to generate strong learners [11]. Gradient Boosting finds problems by using minus-gradient, and betters the performance through computing minus-gradient. Gradient Boosting's goal of each iteration is to reduce the last residuals and build a new model in the Gradient direction with the reduction of residuals. It could be found that the negative gradient of Gradient Boosting Regression (GBR) is the residual, so the thing to fit for the regression problem is the residual.

The steps for implementing GBR are as follows: (1) Initialize the weak learner; (2) Calculate the negative gradient on the sample; (3) Fitting regression trees; (4) Calculate the best fitting values for the leaf region; (5) Update the strong learner; (6) Obtain the expression of the strong learner.

### 2.6 Extreme Gradient Boosting

The eXtreme Gradient Boosting (XGBoost) algorithm is an integrated learning method based on CART decision tree model [12]. Multiple CART decision trees are constructed to provide the accuracy of the forecasting model. Eventually, final predicted value is obtained by summation of the predicted results of the decision tree obtained from each round of training. Compared with GBR algorithm, XGBoost algorithm can effectively prevent overfitting by adding regular terms to the objective function.

During the feature node selection, the values of all feature variables in the training set were traversed, the objective function values before node splitting were subtracted from the sum of the objective function values of the two leaf nodes after node splitting, and the gain value was calculated to obtain the optimal segmentation point of the tree model. The gain value was calculated as follows:

$$L_{\text{split}} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_{\mathrm{L}}} g_i\right)^2}{\sum_{i\in I_{\mathrm{L}}} h_i + \lambda} + \frac{\left(\sum_{i\in I_{\mathrm{R}}} g_i\right)^2}{\sum_{i\in I_{\mathrm{R}}} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \tag{5}$$

## 3 Model Construction Based on Component Data

The main workflow of the oil viscosity determination model is described below. Firstly, input parameters of the machine learning model need to be determined. A lot of elements may influence the oil viscosity, and such influencing factors mainly include two types. One is the composition of crude oil itself, that is, the component data; Second are the external physical factors, such as temperature. Fully including above items during the establishment process of the machine learning model can make the results of the determination of crude oil viscosity more accurate. Therefore, in this study, the data including the component content is used to construct the input parameter dataset of the machine learning model. Attributes and their related ranges employed in input dataset are shown in Table 1.

**Table 1.** Attributes and ranges within input dataset

| Parameter | Unit | Lower limit value | Upper limit value |
|---|---|---|---|
| API at 15 °C | ° | 13.4 | 20.4 |
| %C4 | % | 0 | 1.05 |
| %C5 | % | 0.018 | 2.57 |
| %C6 | % | 0.029 | 3.94 |
| %C7 + | % | 0.884 | 99.52 |
| MWC7 + | g/mol | 295.7 | 418.224 |
| Density at 20°C | g/cm$^3$ | 0.928 | 0.975 |
| Temperature | °C | 20 | 80 |

The oil viscosity forms the output dataset of the machine learning model, which ranges from 14-1431cP. The overall dataset consists of 243 data points. Next, the training dataset and testing dataset are split from whole set in an 8-to-2 ratio. It is worth noting that the input data should be normalized before building the machine learning model.

In this study, six different machine learning approaches which mentioned in Section are employed to construct determination model. In the process of construction, we obtain better prediction performance through appropriate tuning of various machine learning models. Finally, we found the right combination of hyperparameters for the forecasting model based on each machine learning algorithm. For RF model, the number of estimators is 80, while the minimum number of samples required for segmentation is 2. For ET model, the number of estimators is 99, while the minimum number of samples required for segmentation is 2. For DT model, the minimum number of samples required for segmentation is 2, while the maximum number of features in each decision tree is equal to the root of the number of features. For GBR model, the number of estimators is 120, and subsample of GBR is 0.8, while the loss function adopts the mean square error loss function. For SVR model, the value of penalty factor equals to 947.25, while the value of width parameter equals to 126.87. Finally, for XGBoost model, the learning objective function of the model adopts gamma regression, while the value of random_state is set as 1898.

## 4 Results and Discussions

In order to evaluate the performance and the generalization ability of different models, we use the mean value of absolute error (*MAE*) to quantify the comparison. *MAE* of training dataset, testing dataset and overall dataset which obtained by various machine learning models are recorded (Table 2). We could observe the following conclusions from Table 2 that XGBoost model has lowest *MAE* in testing and the whole dataset. In other words, the combination of component data and XGBoost algorithm has higher prediction accuracy than other model combinations.

**Table 2.** *MAE* calculated by vaiours machine learning models

| Algorithm | *MAE* (%) | | |
|---|---|---|---|
| | Training dataset | Testing dataset | Whole dataset |
| RF | 7.17 | 18.92 | 9.06 |
| ET | 2.39E-06 | 19.18 | 3.08 |
| DT | 2.39E-06 | 20.88 | 3.35 |
| GBR | 10.01 | 10.82 | 10.14 |
| SVR | 18.22 | 25.51 | 19.40 |
| XGBoost | 0.24 | 3.89 | 0.82 |

The *MAE* of training dataset calculated by XGBoost model is 0.24%, while the *MAE* of testing dataset calculated by XGBoost model is 3.89%. And he *MAE* of overall dataset calculated by XGBoost model is 0.82%. The error of XGBoost model is very small. All the errors are within the allowable range of engineering calculation and research.

Demonstration of oil viscosity determination results based on composition data and XGBoost model is depicted in Fig. 3. It could be observed from Fig. 3 that all data points are evenly distributed near 45° line, so the prediction accuracy is high. Compared with the ET model (as shown in Fig. 4) with better prediction effect within other methods, the prediction results of testing dataset of the ET model have a large deviation, which proves from the side that XGBoost model has stronger generalization ability compared with other machine learning models applied in this paper.

Compared with conventional determination methods, the oil viscosity determination method combining component data and XGBoost machine learning algorithm has its own advantage which could reduce time and labor cost while ensuring high precision. Through proposed determination approach, the viscosity of crude oil can be quickly and accurately determined.

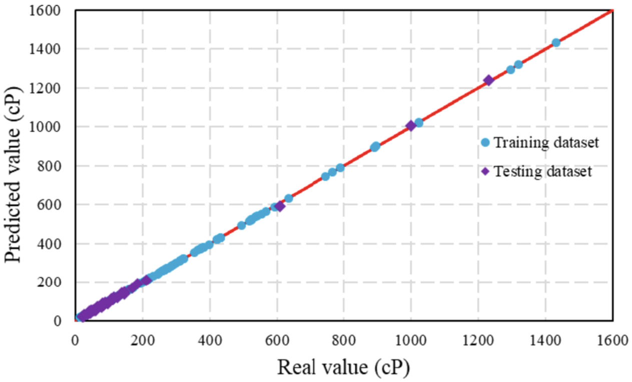


**Fig. 3.** Demonstration of oil viscosity determination results based on composition data and XGBoost model

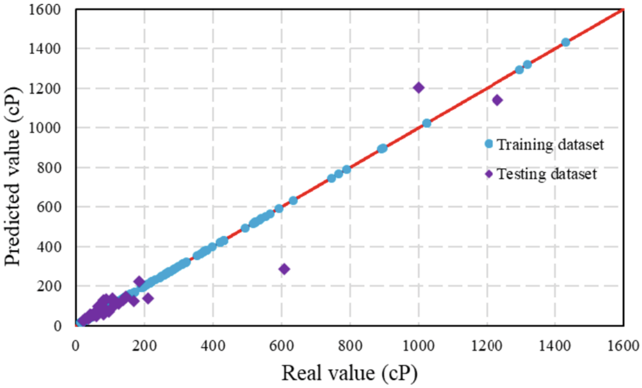


**Fig. 4.** Demonstration of oil viscosity determination results based on composition data and ET model

## 5 Conclusions

(1) In this paper, on the basis of fully considering the factors affecting oil viscosity, eight parameters including composition data were determined as the input parameters of the model, and the oil viscosity value was used as the output parameter to carry out the research.

(2) A series of machine learning algorithm, which including the gradient boosting regression and eXtreme gradient boosting, are used to establish machine learning based model, as for determining oil viscosity. Forecasting performance is compared through calculating error index: mean value of absolute error (*MAE*). The *MAE* of training dataset calculated by XGBoost model is 0.24%, while the *MAE* of testing dataset calculated by XGBoost model is 3.89%. And the *MAE* of overall dataset calculated by XGBoost model is 0.82%. Above research indicated that XGBoost model has great forecasting performance and stronger generalization ability.

(3) Compared with conventional determination methods, the oil viscosity determination method combining component data and XGBoost machine learning algorithm has its own advantage which could reduce time and labor cost while ensuring high precision. Through proposed determination approach, the viscosity of crude oil can be quickly and accurately determined.

(4) Although the data samples used to build the machine learning model in this study are limited, it still demonstrates good predictive performance within the range of viscosity involved. In the future, expanding the dataset can further improve the applicability of model. Moreover, this method can provide some references for similar problems.

**Acknowledgments.** The research is supported by the Major Science and Technology Project of PetroChina (2023ZZ19).

## References

1. Sandor, M., Cheng, Y., Chen, S.: Improved correlations for heavy-oil viscosity prediction with NMR. J. Petrol. Sci. Eng. **147**, 416–426 (2016)
2. Beloglazov, I., Morenov, V., Leusheva, E.: Flow modeling of high-viscosity fluids in pipeline infrastructure of oil and gas enterprises. Egypt. J. Pet. **30**(4), 43–51 (2021)
3. Xu, J., et al.: Insights into the mechanism during viscosity reduction process of heavy oil through molecule simulation. Fuel **310**, 122270 (2022)
4. Wang, H., He, Y., Guo, X.: Viscosity measure and prediction of the $CO_2$-bearing oil system. Petrol. Geol. Recov. Effic. **16**(3), 82–84 (2009)
5. Zhou, X.: Viscosity correlations for gas-free, saturated and under-saturated crude oils. Pet. Explor. Dev. **1**, 44–47 (1991)
6. Sinha, U., Dindoruk, B., Soliman, M.: Machine learning augmented dead oil viscosity model for all oil types. J. Petrol. Sci. Eng. **195**, 107603 (2020)
7. Shi, M., Hu, W., Li, M., Zhang, J., Song, X., Sun, W.: Ensemble regression based on polynomial regression-based decision tree and its application in the in-situ data of tunnel boring machine. Mech. Syst. Signal Process. **188**, 110022 (2023)
8. Cattani, G.: Combining data envelopment analysis and Random Forest for selecting optimal locations of solar PV plants. Energy AI **11**, 100222 (2023)
9. Sachdeva, S., Kumar, B.: Flood susceptibility mapping using extremely randomized trees for Assam 2020 floods. Eco. Inf. **67**, 101498 (2022)
10. Chen, M., Bai, X., Zheng, W., Zhao, D., Wang, Z.: Development index prediction of early polymer flooding based on support vector machine. Fault-Block Oil Gas Field **19**(2), 199–202 (2012)
11. Song, Y., Zhou, H., Wang, P., Yang, M.: Prediction of clathrate hydrate phase equilibria using gradient boosted regression trees and deep neural networks. J. Chem. Thermodyn. **135**, 86–96 (2019)
12. Dong, Y., et al.: A data-driven model for predicting initial productivity of offshore directional well based on the physical constrained eXtreme gradient boosting (XGBoost) trees. J. Petrol. Sci. Eng. **211**, 110176 (2022)

# Practice and Exploration of Data Governance for Drilling Completion

Ling-zhi Jing(✉), Meng Cui, Xin-yi Yang, Yu-meng Tian, and Xiao-yan Shi

CNPC Engineering Technology R&D Company Limited, Beijing, China
jinglingzhi@cnpc.com.cn

**Abstract.** Drilling and completion is the core of energy production and plays an irreplaceable role in improving oil and gas exploration and development and increasing crude oil recovery. Similarly, drilling data is of great significance in drilling optimization, perforation, hydraulic fracturing, and oil and gas production. However, in actual production, data quality and application often fall short, making it difficult to further enhance and optimize oilfield engineering technology. Therefore, effective data governance to address data quality and application problems has become an urgent and critical issue. Additionally, the scale, quality, complexity, and security of data are essential issues that must be considered in data governance processes. To address these issues, this paper proposes an engineering technology data governance approach covering data quality control, data standardization, data modeling, and data mining. These methods and tools have been applied in production practices, and have achieved good results. Additionally, this paper explores the application of intelligent data governance, which aims to quickly and efficiently manage data. This paper compares the advantages and disadvantages of existing data governance algorithms in data governance, providing reference application scenarios for more efficient, reliable, and accurate drilling data governance. In summary, the proposed data governance approach and techniques provide the foundation for further improvements in oilfield engineering technology and management, as well as enhancing data utilization and decision-making for the petroleum service industry. Effective methods and tools must be used to address the challenges of data governance, including data scale, quality, complexity, and security. The proposed data governance approach for engineering

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 434–439, 2024.
https://doi.org/10.1007/978-981-97-0272-5_37

technology, along with its applications and exploration in intelligent data governance, provides significant support for the exploration and implementation of data governance in the petroleum industry.

**Keywords:** Drilling Completion Data · Data Governance · Data Quality Control · Data Standardization · Data Modeling

## 1 Introduction

With the growth of the world economy and population, the demand for energy resources has also increased. Oilfield service engineering technology plays an irreplaceable role in improving oil and gas exploration and development and increasing crude oil recovery. However, oilfield service engineering technology involves a large amount of data, with different types, forms, and varying quality [1]. This has made data management more challenging and calls for more effective data governance approaches.

Data governance refers to the process of managing the lifecycle of data. It involves the standardization and control of data flow, access, update, storage, and deletion processes, ensuring that data is efficient, reliable, secure, and compliant. This, in turn, provides crucial support for enterprise decision-making and business development [2].

Therefore, in order to effectively manage oilfield service engineering technology data, this paper proposes a data governance solution. It focuses on data quality control, data standardization, data modeling analysis, and data mining to achieve unified management and standardized application of data, thereby enhancing the value and utilization efficiency of data. Additionally, this paper explores the application scenarios of commonly used machine learning algorithms in the governance of drilling and well data.

## 2 Overall Architecture

The data governance of engineering technology involves three aspects: the standard system, governance platform, and data applications. Data standardization is an important component of the standard system. Data quality control and data modeling serve as tools in the governance platform, while data mining strengthens data applications and leverages data value (see Fig. 1).

## 3 Specific Implementation Content

### 3.1 Data Quality Control

Data quality refers to whether the data meets the user's requirements and expectations during the processes of data collection, storage, usage, and transmission. It mainly includes aspects such as accuracy, completeness, reliability, and timeliness of the data. In order to ensure the quality of oilfield service engineering technology data, various measures need to be taken for data quality control, as outlined below:

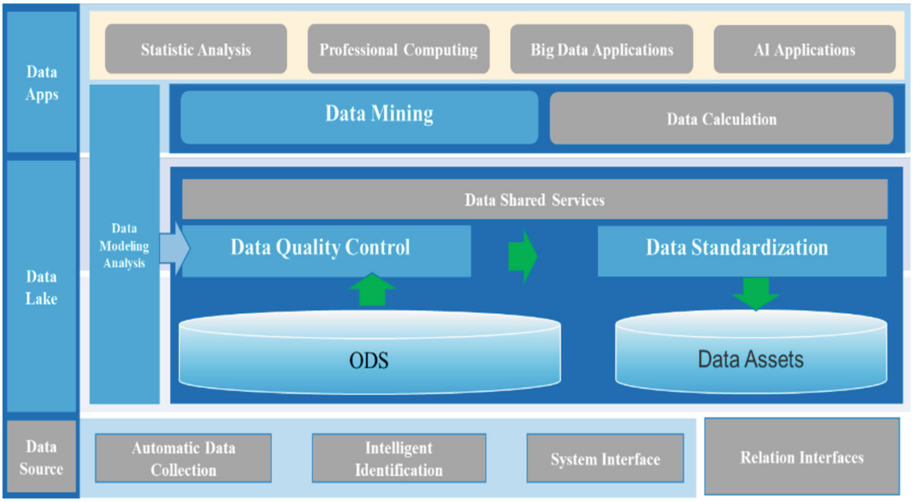


**Fig. 1.** Overall Architecture of Engineering Technology Data Governance.

(1) Data collection control: During the data collection process, data collection standards and rules are established to ensure the correctness and completeness of the data [3].
(2) Data processing control: Data is cleansed, transformed, and subjected to quality analysis to ensure the validity, reliability, and completeness of the data.
(3) Data encryption control: During data transmission, encryption methods are employed to ensure the security of data transmission.
(4) Data backup control: For data backup and recovery tasks, regular backups and testing are conducted to ensure the integrity and reliability of backup data.

### 3.2 Data Standardization

Data standardization refers to the process of integrating and unifying data according to the same semantics, identifiers, and data types. It provides a unified platform for data management, making data storage and sharing more convenient. To achieve the goal of data standardization, the following measures need to be taken:

(1) Establish data standards: Define data storage formats, data definitions, data classification, data naming rules, and other standards to ensure data consistency and adherence to standards [4].
(2) Uniform data formats: Convert and unify data from different sources into a consistent format to facilitate data management and analysis.
(3) Data classification system: Establish a scientific data classification system based on data attributes and characteristics to unify data management and usage [5].
(4) Define data quality assessment indicators: Evaluate and monitor the quality of data, including factors such as data duplicity and data consistency [6].
(5) Establish data security regulations: Clearly define access and usage permissions and regulations for data to ensure data security and confidentiality.

### 3.3 Data Modeling Analysis

Data modeling analysis refers to the process of analyzing and identifying data through the establishment of data models to improve the efficiency and accuracy of data analysis. To achieve data modeling for oilfield service engineering technology data, the following steps need to be taken:

(1) Determine data modeling objectives: Analyze the characteristics and properties of the data to determine the objectives and methods of data modeling [7].
(2) Establish data modeling models: Based on the determined objectives and methods, utilize appropriate data modeling tools and algorithms for data modeling.
(3) Model validation: Improve the accuracy and efficiency of data feature identification and data analysis through data validation and optimization of the model.

### 3.4 Data Mining

Data mining is a method of discovering patterns and knowledge hidden in data through data exploration and analysis, which helps businesses make decisions and manage their operations. To achieve data mining for oilfield service engineering technology data, the following steps need to be taken:

(1) Determine data mining objectives: Analyze business requirements and determine the objectives and methods of data mining.
(2) Search for data patterns: Utilize data mining tools and algorithms to search for and discover hidden patterns and knowledge in the data.
(3) Data analysis and optimization: Improve the efficiency and accuracy of data mining through data analysis and optimization [8].
(4) Apply data mining results: Apply the results of data mining to practical business operations, assisting in decision-making and business management.

## 4 Exploration of Intelligent Data Governance

Artificial intelligence technology can bring higher levels of automation, intelligence, and precision to drilling data governance. Techniques such as machine learning can be utilized to establish predictive models for drilling data, improving data accuracy and prediction capabilities. This article explores the commonly used machine learning algorithms in data governance and summarizes the corresponding application scenarios for each algorithm in drilling data governance.

In drilling data governance, the commonly used machine learning algorithms can be categorized and explored based on their characteristics and application scenarios. The following is a summary of the application scenarios for each machine learning algorithm in drilling data governance:

(1) Regression algorithms: Regression algorithms are widely used in drilling data prediction and modeling [9]. For example, with regression algorithms, prediction models for well trajectories can be established to forecast the variation trends of well trajectories under different conditions, aiding in optimizing drilling plans.

(2) Classification algorithms: Classification algorithms can be applied to classify and recognize drilling data. For instance, classification algorithms can be used to classify drill bit wear levels and determine whether a drill bit replacement is needed, providing suitable recommendations accordingly.
(3) Clustering algorithms: Clustering algorithms help identify similar patterns or clusters within drilling data [10]. For example, by applying clustering algorithms, well trajectory data can be grouped into clusters, finding similar trajectory patterns that can contribute insights and guidance for decision-making [6].
(4) Association rule algorithms: Association rule algorithms are used to mine associations and dependencies within data. In drilling data governance, association rule algorithms can be utilized to discover relationships between variables, uncovering hidden patterns and associations that support data analysis and decision-making.
(5) Reinforcement learning algorithms: Reinforcement learning algorithms are useful for optimization and decision-making problems in drilling data governance. For example, in optimizing wellbore trajectories, reinforcement learning algorithms can be used to gradually determine the optimal operational strategies through interactions with the environment, achieving automation and intelligence in well trajectory design.

In addition to the mentioned algorithms, there are other machine learning algorithms that can be applied to drilling data governance, such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Decision Trees [11]. Each algorithm has its own unique characteristics and suitable range of applications. The choice of algorithm depends on the specific drilling data and governance needs.

## 5 Conclusion

For the governance of oilfield engineering technical data, it is essential to manage and utilize the data from aspects such as data quality control, data standardization, data modeling, and data mining. Machine learning algorithms can be combined with drilling data governance to achieve automation, intelligence, and precision in drilling data governance, as needed and appropriate. Only through unified management and standardized application of oilfield engineering technical data can the value and effectiveness of the data be maximized. Only by fully integrating with new technologies can the development and innovation of oilfield engineering technical data governance be promoted. Therefore, we recommend that relevant companies prioritize the work of oilfield engineering technical data governance and establish comprehensive data governance mechanisms and standards to contribute to the sustainable development of data.

## References

1. Zhu, F.Y.: Construction of an ocean petroleum drilling and completion data analysis system based on big data environment. Adhesion **49**(05), 117–121 (2022)
2. Ren, F.S.: Construction and statistical application research on drilling and cementing fluid data governance system in well completion. Petrochem. Technol. **29**(02), 113–116 (2022)

3. Zhang, F.X.: Research on data governance engineering and application technology of intelligent oilfield. Inf. Syst. Eng. **353**(05), 52–54 (2023)
4. Yang, F.L., Wan, S.J., Qiu, T.Q.: Research on data governance engineering and application technology of intelligent oilfield. Inf. Syst. Eng. **341**(05), 149–152 (2022)
5. Sun, F.E., Zu, S.Z.: Data governance engineering and its application in intelligent oilfields. China Inf. **331**(11), 78–79 (2021)
6. Li, F.Y.: Application of artificial intelligence in data governance. Inf. Commun. Technol. Policy **299**(05), 23–27 (2019)
7. Xiang, F.M., Li, S.Y., Xin, T.L.: Research on data governance and application technology of intelligent oilfields. China Manag. Inf. **24**(22), 82–83 (2021)
8. Liang, F.Z.: Urgent need to build a reasonable and efficient data governance system in the era of artificial intelligence. Natl. Gover. **295**(31), 40–42 (2020). https://doi.org/10.16619/j.cnki.cn10-1264/d.2020.31.008
9. Li, F.L.: Application research on data intelligence classification technology in data governance. Satellite TV Broadband Multimedia **514**(09), 15–17 (2020)
10. Huang, H., Yan, F.: Research on data governance engineering and application technology of intelligent oilfields. China Manag. Inf. **23**(06), 68–69 (2020)
11. An, F.P., Yan, S.J.: Intelligent data governance platform. In: China Cybersecurity Industry Alliance, China Electronics Technology Standardization Research Institute (eds.) Proceedings of the National Cybersecurity Awareness Week “Cybersecurity Industry Development Forum”. Information Security Research Journal, no. 4 (2021). https://doi.org/10.26914/c.cnkihy.2021.038945

# Ontology Construction Technology of Knowledge Graph in Oil and Gas Exploration and Development

Ning Li[1,2](✉), Liang Ren[1,2], Zong-shang Liu[1,2], Shu-hang Ren[1,2], Chong Xiang[1,2], Bo-yu Wu[1,2], and Xuan Cai[1,2]

[1] Research Institute of Petroleum Exploration and Development, PetroChina, Beijing, China
lining_riped@petrochina.com.cn

[2] Artificial Intelligence Technology R&D Center for Exploration and Development, CNPC, Beijing 100083, China

**Abstract.** With the gradual application of knowledge atlas technology in oil and gas exploration and development fields such as intelligent evaluation of oil and gas reservoir and intelligent identification of oil and gas reservoir through logging, the application of knowledge Atlas in oil and gas industry has attracted more and more attention. In view of the problem of knowledge expression and application in E&P sector field, a professional word bank of E&P was established based on web crawler technology, a feature corpus of E&P field was established by extracting E&P sector related achievement documents and literature data, and the principle of E&P knowledge graph ontology construction was formulated and the construction process was defined. The multistage classification system of oil and gas exploration and development knowledge atlas ontology is described and constructed from multiple dimensions such as sector, object and feature, and the knowledge atlas ontology model of oil and gas exploration and development sector domain is formed, which lays a foundation for the in-depth application of knowledge atlas technology in petroleum exploration and development.

**Keywords:** Oil and gas exploration and development · Knowledge graph · Ontology · Relationship · Object · Features

## 1 Introduction

With the continuous progress of technology and the continuous development of exploration and development technology in the petroleum industry, a huge amount of sector knowledge and experience has been accumulated, which is an important basis for exploration and development decision-making. How to accurately push these knowledge and experiences to the researchers in need is an important issue in the construction and application of knowledge graphs.

The knowledge graphs was first proposed as the concept of Semantic Web in 2006, and formally proposed by Google in 2012. Its purpose is to provide users in all industries with accurate and intelligent search experience. At present, the knowledge graph [1, 2]

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 440–460, 2024.
https://doi.org/10.1007/978-981-97-0272-5_38

has been widely applied by multiple industries in fields such as intelligent question answering and intelligent recommendation.

The construction of knowledge graph ontology is the core and foundation of knowledge graph construction, and ontology construction needs to be combined with specific application fields. Among them, the construction of knowledge graph ontology for exploration and development sector needs to build a scientific and complete knowledge classification system based on comprehensive consideration of the current knowledge situation in the exploration and development field, use the ontology construction language to express the knowledge classification system clearly and clearly, and establish the exploration and development domain knowledge graph ontology.

The construction of oil and gas industry knowledge graph [3–5] is still in its infancy. At present, under the background of inferior quality of oil and gas resources, diversification of exploration objectives, complex development objects, and fierce market competition, conventional on-shore oil and gas exploration technology is not adaptable, water drive oilfield development efficiency of high water cut oilfield is low, energy consumption is high, water cut is rising fast, low natural productivity of oil Wells developed in low permeability/ultra-low permeability oilfield, low production of single well, high investment of production capacity of million tons. China's petroleum industry is faced with many technical challenges in domestic and overseas oil and gas exploration and development. Artificial intelligence and knowledge graph technology can play a revolutionary role beyond imagination at critical moments, but to realize the deep application of knowledge graph technology in our oil industry, There are many technical difficulties in the field of oil and gas exploration and development, such as the lack of comprehensive knowledge graph [6–8] system, the absence of intelligent analysis ontology and entity model, the lack of supporting intelligent application technology and the imperfection of supporting platform functions. Therefore, the current research focus is mainly on the innovative construction of oil and gas exploration and development knowledge ontology.

In this study, the ontology construction closely focuses on the research task of exploration and development knowledge graph construction, combines the current status and needs of exploration and development sector, closely combines the knowledge graph ontology construction with core sector, conducts research and analysis on the status quo of PetroChina's exploration and development sector [9], clarifies the scope of knowledge graph ontology construction, determines the construction content of exploration and development knowledge system, and formulates sector classification and sector object sorting strategy, Guide the construction of the ontology of exploration and development knowledge graph.

In terms of ontology construction foundation: Based on the petroleum encyclopedia of the Petroleum Industry Press, petroleum reference materials, and books, establish a professional vocabulary and special corpus in the field of petroleum exploration and development, as the foundation for the ontology construction of the exploration and development knowledge graph.

In terms of ontology construction principles, it takes sector as the main line and objects and features as the auxiliary for multi-dimensional division. On the basis of complete sector description, clear structure and description specification, optimization is carried out by referring to PetroChina exploration and development knowledge graph, artificial intelligence platform technical specification and EPDM model, and combining with the correlation relationship between sector attributes and knowledge.

In terms of ontology construction process: Following the theoretical standards of top-level design for exploration and development of PetroChina, the top-level concepts of discipline and industry standards are combined with actual production application data concepts, and the sector ontology in the exploration and development field is refined and improved at a hierarchical level.

Based on the above thinking scheme, this paper constructs the ontology of domain knowledge graph of exploration and development. The purpose of the construction is to solve the technical problems of representation, sharing and application of domain knowledge of exploration and development, and truly apply the knowledge graph to oil and gas production.

## 2 Workflow

According to the characteristics of domain knowledge of oil exploration and development, this paper classifies the knowledge, uses the Scrapy web crawler technology to obtain various knowledge words from the Internet pages, forms a professional vocabulary of exploration and development, and collects and sorts a large number of exploration and development literature data to form a characteristic corpus of various types of knowledge.

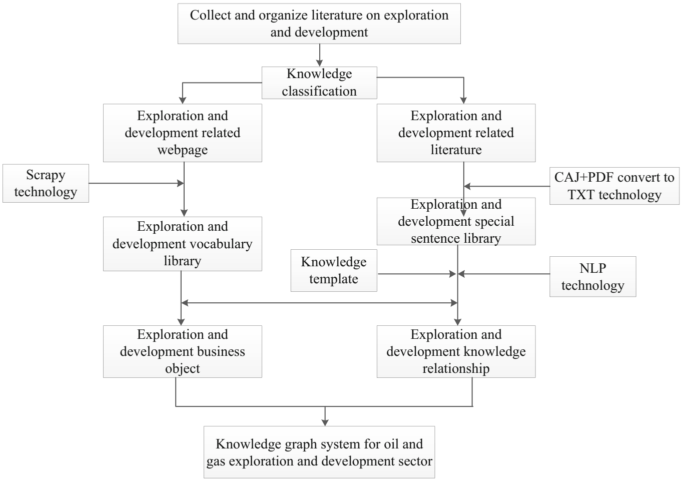

Collect and organize literature on exploration and development
Knowledge classification
Exploration and development related webpage
Exploration and development related literature
Scrapy technology
CAJ+PDF convert to TXT technology
Exploration and development special sentence library
Exploration and development vocabulary library
Knowledge template
NLP technology
Exploration and development business object
Exploration and development knowledge relationship
Knowledge graph system for oil and gas exploration and development sector

**Fig. 1.** Workflow for constructing an ontology of exploration and development knowledge graph.

It uses the knowledge templates of various corpora and NLP technology to extract sector object and knowledge relationships of various types of exploration and development knowledge, Forming an ontology of exploration and development knowledge graph in the field of oil and gas sector. The specific technical roadmap is shown in Fig. 1.

# 3 Construction Ideas for the Ontology of Exploration and Development Knowledge Graph

The construction of knowledge graph ontology of exploration and development closely focuses on exploration and development research tasks, combines the status quo and needs of exploration and development sector, closely combines the construction of knowledge graph ontology with core sector, conducts research and analysis on the status quo of PetroChina's exploration and development sector, defines the construction scope of knowledge graph ontology, determines the construction content of exploration and development knowledge graph ontology, and formulates sector classification and sector object sorting strategy, Guide the construction of the ontology of exploration and development knowledge graph.

## 3.1 Foundation for Constructing Ontology of Exploration and Development Knowledge Graph

The construction of a knowledge graph ontology for oil and gas exploration and development is based on the establishment of a professional vocabulary and feature corpus for oil and gas exploration and development.

**Professional Vocabulary for Oil and Gas Exploration and Development.** Based on the core sector knowledge of oil and gas exploration and development, establish a professional vocabulary library in the field of oil and gas exploration and development, covering drilling, geological exploration, logging, geophysical exploration, oil and gas field development and production, oil and gas gathering and transportation, etc., to ensure coverage of profession-al vocabulary involved in oil and gas exploration and development.

The vocabulary of the professional vocabulary comes from the Petroleum Encyclopedia column of the Petroleum Industry Press, which provides explanations of professional vocabulary and corresponding terms for each section of the petroleum industry. In order to collect vocabulary in the field of exploration and development, Scrapy web crawler technology is used to crawl all vocabulary in the exploration block. Scrapy is a program and script that automatically retrieves web page information according to certain rules. Different from traditional search engines, web crawler automatically crawl data and only crawl specific types of information they want to get, so they save time and improve the efficiency of search engines [10–12] (Fig. 2).

Based on the principle of Scrap crawling, the vocabulary crawling in the petroleum encyclopedia was implemented using Python language. The detailed process is as follows:

(1) Obtain user-defined parameters and the URL of the Petroleum Press;
(2) Jump to the homepage of Petroleum Press;
(3) Obtain the menu bar location of the Petroleum Encyclopedia and enter this interface;
(4) Crawl the relevant vocabulary on the Petroleum Encyclopedia page;
(5) Store the crawled entry data into an excel document;

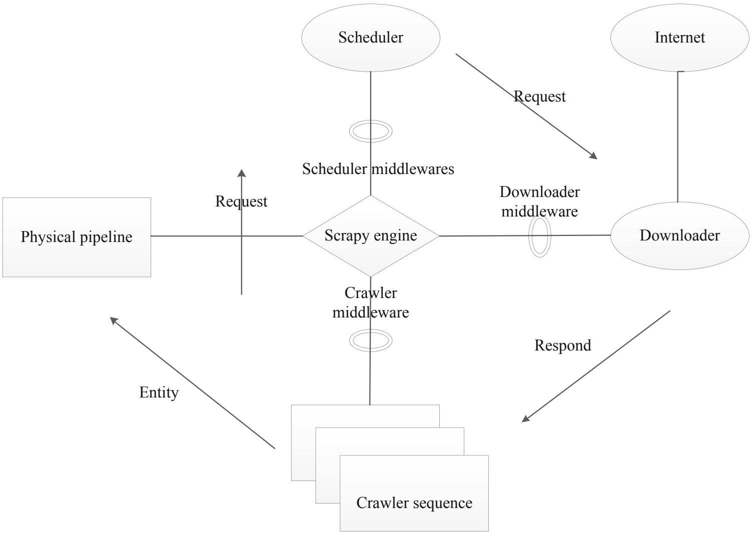


**Fig. 2.** Workflow of the Scrapy framework.

(6) Determine whether the next entry exists. If it does, proceed to the next page and continue crawling to the next entry. If it does not exist, export the final Excel document;
(7) End of crawling.

After crawling, a total of 6135 professional terms were obtained, including geological exploration, geophysical exploration, logging, oil and gas field development and extraction, drilling, and oil and gas gathering and transportation (as shown in Table 1), which basically meet the naming entity requirements for constructing an exploration knowledge graph (Fig. 3 and Table 2).

**Oil and Gas Exploration and Development Feature Library.** Establish a feature corpus for the field of oil and gas exploration and development by extracting documents and literature related to exploration and development sector [13, 14]. The feature corpus includes corpus related to physical and chemical exploration, wellbore engineering, comprehensive research, analysis and testing, oil and gas development and production, and surface engineering sector, and continuously supplements and improves the corpus based on the knowledge graph research stage. Based on the sector and application characteristics of exploration and development, the corpus is divided into "normative class, method class, expert class, and case class" (Fig. 4).

(1) Normative class

Collect and organize various specifications in the field of petroleum exploration and development, store them in the form of documents or charts, and reflect the work templates, standards, and other contents related to the exploration and development field sector, such as technical specifications for data interpretation, drilling and coring quality indicators, etc.

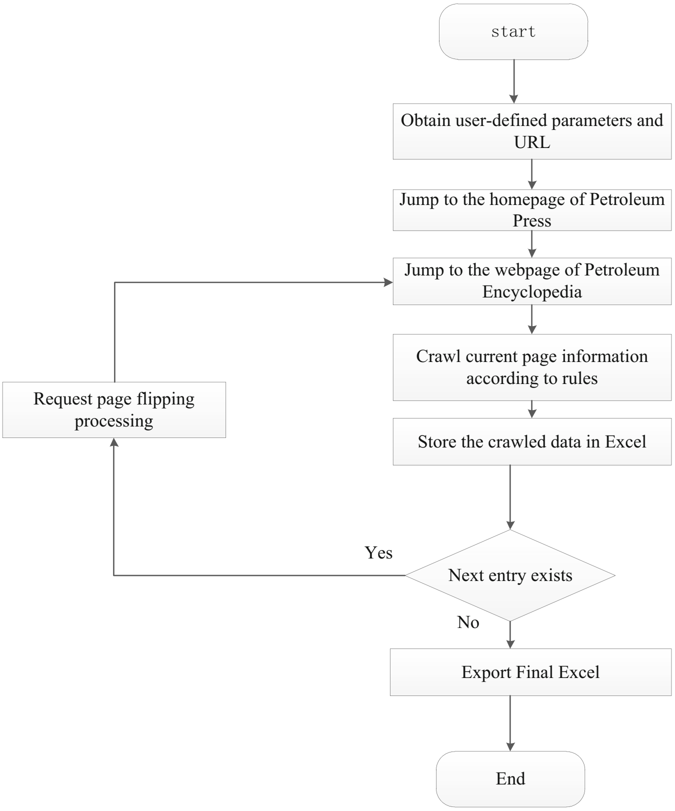

start
Obtain user-defined parameters and URL
Jump to the homepage of Petroleum Press
Jump to the webpage of Petroleum Encyclopedia
Crawl current page information according to rules
Store the crawled data in Excel
Next entry exists
Yes
Request page flipping processing
No
Export Final Excel
End

**Fig. 3.** Crawling professional vocabulary process.

**Table 1.** Comparison between different work areas.

| Oil & gas E&D sector | Number of vocabulary |
|---|---|
| Geological prospecting | 1600 |
| Geophysical exploration | 916 |
| Well logging | 671 |
| Oil and gas field development and extraction | 1190 |
| Well drilling | 1116 |
| Oil and gas gathering and transportation | 642 |
| Total | 6135 |

**Table 2.** Example of Classification of Professional Entries.

| Geological prospecting | Oil Geophysical Prospecting | Well logging | Oil and gas field development | Well drilling engineering | Oil and gas gathering and transportation |
|---|---|---|---|---|---|
| Earth body | Geophysical exploration | Geophysical logging | Reservoir Physics | well drilling | Field oil and gas gathering and transportation |
| Earth mass | Oil Geophysical Prospecting | Cable logging | Reservoir fluid | Drilling design | Oil and gas distribution |
| Earth gravity | Seismic exploration | Mining site Spheric physics | Standard state | Drilling methods | Oil gas mixed transportation |
| Earth radioactivity | Reflected wave method | Logging while drilling | Oilfield water | Drilling method | Gathering and transportation process |
| ground temperature | Refractive wave method | Electrical logging | Coexisting water | Rotary drilling method | Open process |
| Paleogeothermal temperature | Physical properties of rocks | Resistivity logging | Bound water | Portable drilling | Closed process |
| Exothermic layer | Rock elasticity | Ordinary electrode system | Reservoir oil | Developing sand drilling | Can free process |
| Normal temperature layer | Rock wave velocity | Gradient electrode system | Degassed oil | Classification of wells | Oil shop sealed air extraction |
| Earth heat flow | Lame coefficient | Electrode coefficient | Two-phase volume coefficient | Shaw rock hardness | Non oral backpressure |
| Geothermal anomaly | Young's modulus | Electrode system Record points | Reservoir fluid compressibility coefficient | Smith rock hardness | Oil outlet pipe |
| Temperature field | Poisson's ratio | Electrode system Electrode spacing | Gas oil ratio | Mineral elastic modulus | Oil collection pipe |
| Rock thermal conductivity | Bulk modulus | Detection range | Dissolved gas oil ratio | Rock elastic modulus | Radial pipe network |
| … … | … … | … … | … … | … … | … … |

(2) Method class

Collect and organize various work methods in the field of petroleum exploration and development, store them in the form of documents and mathematical formulas, and reflect the work methods, rules, formulas, and literature related to the sector in the

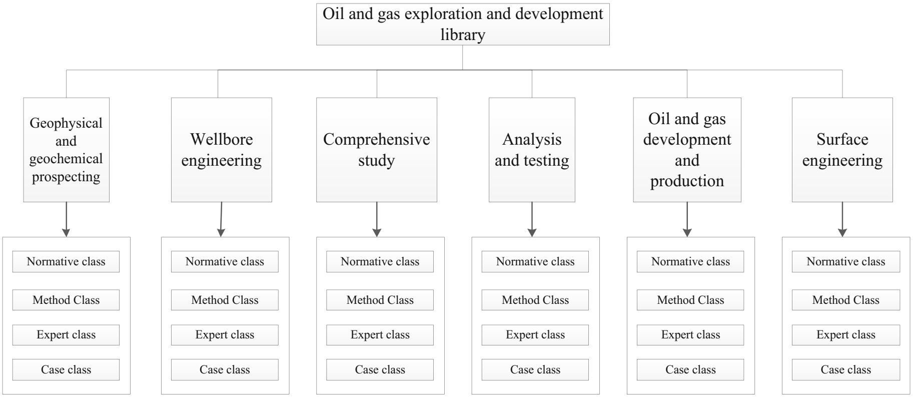


**Fig. 4.** Library of oil and gas exploration and development characteristics.

exploration and development field. Different processes and technical methods are used for sorting out different sectors, such as downhole operations including acidification and sand control, as well as centrifugal and diaphragm methods for measuring capillary pressure curves.

(3) Expert class

This includes personal information, work experience, work achievements, work summaries, and other knowledge of experts in the field of petroleum exploration and development.

(4) Case class

This includes actual sector cases in the field of petroleum exploration and development, stored in the form of documents, reflecting actual project cases, experiences, and achievements related to sector work in the exploration and development field. A variety of cases are generated in the exploration and development sector, such as drilling and completion design cases, development plan preparation cases, etc.. Different cases are sorted out for different sector object.

## 3.2 Principles for Constructing Ontology of Exploration and Development Knowledge Graph

The purpose of designing the knowledge ontology classification framework is to logically partition and organize a large number of entries based on common features, in order to facilitate storage, sharing, and knowledge graph application. For the field of exploration and development, knowledge classification should be guided by petroleum related topics, and the overall classification of the main objects and sectors of the oilfield should be summarized and summarized. Based on the exploration and development sector as the main line, the life cycle of oil and gas field exploration and development, the management stage of oil and gas field exploration and development, and other dimensions as the principles, the core sector knowledge and functional knowledge involved in exploration and development are divided into categories to form a classification plan for exploration and development knowledge, ensuring coverage of all exploration and development sectors.

During the process of constructing the ontology, six aspects were analyzed and considered (Table 3).

**Table 3.** Classification of oil and gas field exploration and development sector.

| Sector Area | Sector domain | Number of vocabulary |
|---|---|---|
| Oil & gas exploration and development | Physical and chemical exploration | 1600 |
| | Wellbore engineering | 916 |
| | Comprehensive research | 671 |
| | Analysis and testing | 1190 |
| | Oil and gas development and production | 1116 |
| | Surface engineering | 642 |

The main dimensions that can be adopted for the classification of exploration and development knowledge are:

(1) Divided by sector and sector process;
(2) By sector object;
(3) Divided by professional (characteristic) knowledge.

The design of exploration and development knowledge classification follows the principles of "complete sector, clear structure, standardized sector description, and good scalability", and is optimized based on the EPDM model. The main dimension is "sector object", forming the ontology classification of exploration and development knowledge. On the basis of the knowledge classification system, feature descriptions are provided for the objects involved in the knowledge content, providing a basis for the application of knowledge.

Taking the sector as the main line, supplemented by objects and features for multidimensional division.

The sector division mainly refers to the knowledge graph of China Petroleum Exploration and Development, the technical specifications of artificial intelligence platforms, and the EPDM model, and is optimized based on the correlation between sector attributes and knowledge. The optimization of sector is mainly based on the principle of meeting the fine-grained requirements of sector for knowledge application, effectively distinguishing knowledge, and facilitating knowledge application.

(1) Sector integrity. Full coverage of exploration and development sector.
(2) Clear sector structure and unique sector classification.
(3) Standardized sector description and accurate sector division description.

(4) The granularity of sector division should be greater than or equal to the granularity of knowledge, and a complete knowledge should not be divided into content due to the subdivision of sector classification.
(5) Sector classification has good scalability. Fully consider reserving new sector. Ensure the continuous development of exploration and development sector, while ensuring the stability of sector classification.

Through multiple investigations, including exploration and development knowledge graph and artificial intelligence platform technical specifications, EPDM models, textbooks, enterprise standards, Sinopec system, and a large number of related literature, the first and second level knowledge classification of oil and gas exploration and development knowledge graph has been determined. Subsequently, based on detailed research and expert communication, a three-level and four-level knowledge classification of the oil and gas exploration and development knowledge graph was determined (Fig. 5).

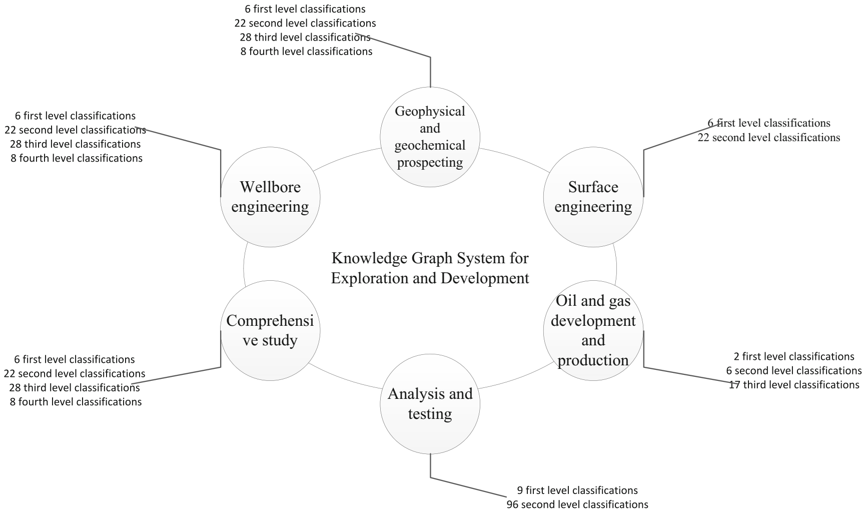

6 first level classifications
22 second level classifications
28 third level classifications
8 fourth level classifications
Geophysical and geochemical prospecting
6 first level classifications
22 second level classifications
28 third level classifications
8 fourth level classifications
Wellbore engineering
6 first level classifications
22 second level classifications
Surface engineering
Knowledge Graph System for Exploration and Development
Comprehensi ve study
Oil and gas development and production
6 first level classifications
22 second level classifications
28 third level classifications
8 fourth level classifications
2 first level classifications
6 second level classifications
17 third level classifications
Analysis and testing
9 first level classifications
96 second level classifications

**Fig. 5.** Exploration and development knowledge graph classification diagram.

### 3.3 Construction Process of Knowledge Graph Ontology for Oil and Gas Exploration and Development

On the basis of the constructed knowledge classification system, based on the exploration and development knowledge graph and artificial intelligence platform technical specifications, the ontology construction technology is used to express the knowledge classification system clearly, and to establish the exploration and development domain ontology.

The ontology construction of exploration and development knowledge graph [15, 16] is a system engineering that involves multiple sector fields such as physical and chemical exploration, wellbore engineering, comprehensive research, oil and gas field development and production, analysis and testing. The amount of sector knowledge is huge and complex among each other. In order to clearly and accurately construct the ontology of exploration and development knowledge graph, the specific construction idea is:

(1) Following the top-level design theory of exploration and development, the ontology planning and design are carried out from top to bottom, with priority given to defining the top-level comprehensive and comprehensive concept expression in the exploration and development sector field. Based on the artificial intelligence platform and knowledge graph technical specifications, referring to the EPDM model, after analysis and sorting, the top-level framework of exploration and development ontology is divided into physical and chemical exploration, wellbore engineering, comprehensive research, analysis and testing, oil and gas development and production surface engineering, based on this concept, is refined step by step, constantly sorting and improving new concepts at different levels.
(2) Improve the construction of exploration and development ontology from bottom to top, by collecting actual data from exploration and development data sources, focusing on practical application research sector, combined with the exploration and development sector entities extracted from actual data, fully considering the application scenarios and user actual needs of the ontology, and logically analyzing and organizing a large number of actual data entries based on common features, At the same time, based on the top-level design of exploration and development, refine the ontology concepts under the sector application [17–20].
(3) Effectively combining top-down and bottom-up methods, combining top-level concepts of discipline and industry standards with actual production application data concepts, and refining and improving the sector ontology of exploration and development at a hierarchical level.

For the construction of ontology, different fields have different understandings and methods, but ontology construction basically follows the following principles:

(a) The principle of objectivity, the definition of ontology follows industry standards and norms;
(b) The principle of completeness, the definition of ontology can fully ex-press the meaning of industry terms;
(c) The principle of consistency is that the conclusions drawn from the inference of knowledge are consistent with the meaning of its ontology;
(d) The principle of continuous improvement is that the exploration and development sector is a gradual evolution and continuous improvement process. The exploration and development sector continues to develop, and due to various reasons such as mechanism adjustments, market changes, and sector expansion, the exploration and development ontology will also change accordingly;
e) The principle of sector optimization, the construction of exploration and development ontology, helps enterprises optimize (sector processes and provide better products and services with higher efficiency and better quality.

The specific construction process is shown in the following figure:

(4) The granularity of sector division should be greater than or equal to the granularity of knowledge, and a complete knowledge should not be divided into content due to the subdivision of sector classification.

(5) Sector classification has good scalability. Fully consider reserving new sector. Ensure the continuous development of exploration and development sector, while ensuring the stability of sector classification.

During the process of constructing the ontology, six aspects were analyzed and considered (Fig. 6).

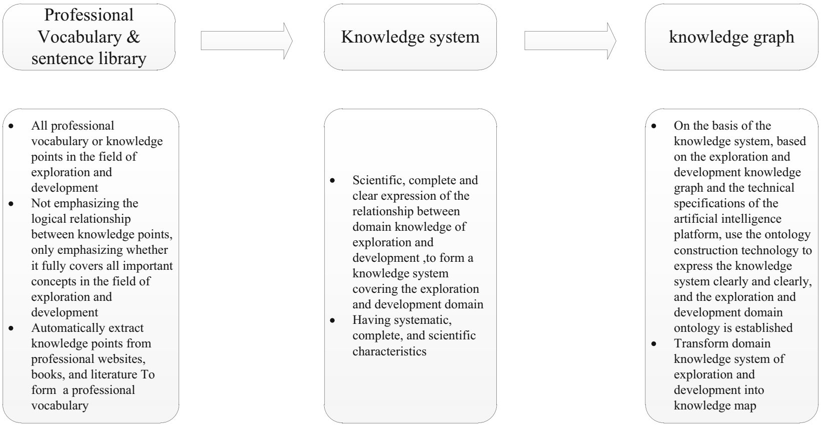

Professional Vocabulary & sentence library
Knowledge system
knowledge graph
• All professional vocabulary or knowledge points in the field of exploration and development
• Not emphasizing the logical relationship between knowledge points, only emphasizing whether it fully covers all important concepts in the field of exploration and development
• Automatically extract knowledge points from professional websites, books, and literature To form a professional vocabulary
• Scientific, complete and clear expression of the relationship between domain knowledge of exploration and development ,to form a knowledge system covering the exploration and development domain
• Having systematic, complete, and scientific characteristics
• On the basis of the knowledge system, based on the exploration and development knowledge graph and the technical specifications of the artificial intelligence platform, use the ontology construction technology to express the knowledge system clearly and clearly, and the exploration and development domain ontology is established
• Transform domain knowledge system of exploration and development into knowledge map

**Fig. 6.** Process of knowledge ontology for oil and gas exploration and development knowledge graph.

# 4 Exploration and Development Knowledge Graph Ontology Model

## 4.1 Construction of Exploration and Development Knowledge Graph Ontology

**Sector Domain.** The sector domain focuses on exploration and development sector, with multiple dimensions such as the life cycle of oil and gas field exploration and development, and the management stage of oil and gas field exploration and development as the principle. The core sector knowledge and functional knowledge involved in exploration and development are divided into categories. In accordance with the habits of oil and gas exploration and development management, the sector within different sector domains should not overlap and ensure coverage of all exploration and development sectors. Based on the above principles and methods, the oil and gas field exploration and development sector is divided into six major sector domains: “physical and chemical exploration”, “wellbore engineering”, “comprehensive research”, “analysis and testing”, “oilfield development and production”, and “surface engineering”.

(1) Physical and chemical exploration

To ensure that the field of geophysical and geochemical exploration does not duplicate and can cover all sectors, sector division is carried out based on actual sector work and scenarios, and a geophysical and geo-chemical exploration ontology model is established. The sector model involved in geophysical and geochemical exploration is established through the top-level sector of geophysical and geochemical exploration set by experts in the petroleum field. Based on multiple characteristics such as the content, object, and scope of the sector, the multi-level sector ontology model architecture is formed, which mainly includes all content related to geophysical and geochemical exploration deployment, seismic exploration, gravity exploration, magnetic exploration, chemical exploration, and electrical exploration (Table 4).

**Table 4.** Examples of Physical and Chemical Exploration Sector Concepts.

| Sector domain | First level sector | Secondary sector | Third level sector |
|---|---|---|---|
| Physical and chemical exploration | Physical and chemical exploration deployment | Seismic exploration deployment | Surface seismic deployment |
| | | | Well site seismic deployment |
| | | Gravity exploration deployment | – |
| | | Magnetic exploration deployment | – |
| | | Chemical exploration deployment | – |
| | | Deployment of electrical exploration | – |

(2) Wellbore engineering

The wellbore engineering constructs a multi-level sector ontology model architecture based on the concept of physical and chemical exploration sector do-main, mainly including logging, logging, drilling, testing, and all other content related to the wellbore (Table 5).

(3) Comprehensive research

Comprehensive research constructs a multi-level sector ontology model architecture based on the concept of physical and chemical exploration sector field, mainly including planning and deployment, exploration geological research, development geological research, oil and gas reservoir engineering, development scheme design, and all other related content related to comprehensive research (Table 6).

**Table 5.** Example of Wellbore Engineering Sector Concept.

| Sector domain | First level sector | Secondary level sector | Third level sector | Fourth level sector |
|---|---|---|---|---|
| Wellbore engineering | Well design | Well location design | – | – |
| | | Well position measurement | – | – |
| | | Drilling geological design | – | – |
| | | Drilling Engineering Design | – | – |
| | Drilling engineering | Drilling construction | Predrilling | – |
| | | | Drilling operations | Normal drilling operations |
| | | | | Directional homework |

**Table 6.** Example of Comprehensive Research Sector Concept.

| Sector domain | First level sector | Secondary sector | Third level sector |
|---|---|---|---|
| Comprehensive research | Planning and Deployment | development planning | Overall development plan |
| | | | Special planning |
| | | Exploration Planning and Deployment | Exploration Planning and Design |
| | | | Exploration Deployment Design |
| | | Development Planning and Deployment | Development planning and design |
| | | | Development Deployment Design |
| | Exploration geological research | Stratigraphic research | Determination of Standard Layer and Sequence Boundary |
| | | | Single well research |

(4) Analysis and testing

The analysis and testing builds a multi-level sector ontology model architecture according to the thinking of geophysical and geochemical exploration sector field, mainly including conventional core analysis, core specific analysis, fluid property analysis, rock mineral analysis, stratum paleontology analysis, geochemical analysis, enhanced oil recovery indoor analysis and all other contents related to analysis and testing (Table 7).

**Table 7.** Example of Analytical and Laboratory Sector Concepts.

| Sector domain | First level sector | Secondary sector |
|---|---|---|
| Experimental analysis | Routine core analysis | Porosity analysis |
| | | Permeability analysis |
| | | Oil water saturation analysis |
| | | Analysis of carbonate content in rocks |
| | | Determination of chloride content in rocks |
| | Special core analysis | Determination of oil-water relative permeability by steady-state method |
| | | Measurement of oil-water relative permeability by unsteady state method |

(5) Oil and gas development and production

The multi-level sector ontology model architecture for oil and gas development and production is constructed based on the concept of physical and chemical exploration sector, mainly including all content related to oil and gas development and production, such as oil and gas reservoir engineering management and oil and gas pro-duction engineering management (Table 8).

(6) Surface engineering

Surface engineering constructs a multi-level sector ontology model architecture based on the concept of geophysical and geochemical exploration sector domain, mainly including surface engineering design, surface engineering construction, and all other related content related to surface engineering (Table 9).

**Object Domain.** Sector object [4] refers to sector related objects in oil and gas exploration and development, including both sector object with spatial location attributes (positioning sector object), such as structural units and wells, and objects without spatial attributes (non positioning sector object), such as documents. According to the six sector domains of geophysical and geochemical exploration, wellbore engineering, comprehensive research, analytical testing, oil and gas development and production, and surface

**Table 8.** Example of oilfield development and production sector concepts.

| Sector domain | First level sector | Secondary sector | Third level sector |
|---|---|---|---|
| Oil & gas development & production | Oil & gas reservoir engineering management | Production plan | Single well production plan |
| | | | Oil and gas field production plan |
| | | Oil and gas production | Single well production |
| | | | Oil and gas field production |
| | | Dynamic analysis | Single well performance analysis |
| | | | Well group dynamic analysis |
| | | | Dynamic analysis of block units |
| | Oil and gas engineering management | Injection and production engineering management | Machine procurement system management |
| | | | Analysis of underground technical conditions |
| | | | Water injection system management |

**Table 9.** Example of Surface Engineering Sector Concept.

| Sector domain | First level sector | Secondary sector |
|---|---|---|
| Surface engineering | Surface engineering design | Electronic power network design |
| | | Pipeline design |
| | | Station storage design |
| | | Oilfield facility design |
| | | Oil production platform design |
| | Surface engineering construction | Pipeline construction |
| | | Station storage construction |
| | | Electronic power network construction |
| | | Oil production platform construction |

engineering involved in exploration and development, and combined with the results of knowledge classification, the relevant objects of oil and gas exploration and development sector are sorted out, basically covering the objects involved in exploration and development sector, and sector object are divided uniformly according to sector types (Table 10).

**Table 10.** Example of concept of exploration and development sector object.

| Basin | – | – | – |
|---|---|---|---|
| – | Depression | | – |
| – | – | Oil and gas fields | – |
| – | – | – | Block |
| Chrono stratigraphy | – | – | – |
| Seismic reflection interface | – | – | – |
| Oil and gas horizons | – | – | – |
| – | Oil reservoir group | – | – |

**Feature Domain.** The description of exploration and development characteristics is based on earth science, including petrology, mineralogy, paleontology, stratigraphy, structural geology and other disciplines. Each discipline knowledge has similarities and differences. To describe the sector object, the following three feature description conditions must be fully considered: (1) discipline knowledge can completely describe the meaning of this feature of the sector object; (2) The disciplinary knowledge describing the characteristics of exploration and development can not only reflect the common characteristics of different features, but also reflect the different characteristics of each feature; (3) Having continuous improvement characteristics, able to continuously optimize and supplement exploration and development feature description knowledge in response to the development and changes of the discipline.

Exploration and development involves a large number of geoscience disciplines. To build a feature de-scription ontology in the exploration and development field, a certain discipline node is required as the top-level concept of feature description. Taking "sedimentary rock" as an example, the feature concept of the discipline knowledge is described in detail to assist ontology construction (Fig. 7).

**Relationship.** Relationship is the description of the relationship between two concepts. After concept extraction and definition, it becomes scattered and isolated individual units. It is necessary to extract and analyze the relationships between concepts, establish the relationship between concepts, and form a network association structure. When constructing the exploration and development ontology, it is necessary to construct the

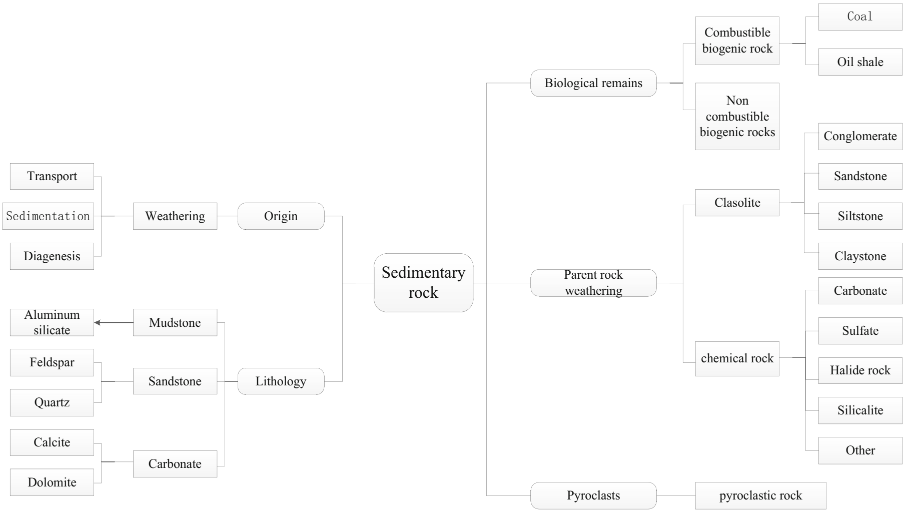


**Fig. 7.** Example of sedimentary rock concept.

relationships between concepts based on the actual situation of the oil and gas sector field. On the basis of general relationships be-tween concepts, this article further establishes ontology custom relationships specifically for concepts in the petroleum field according to sector ontology requirements, and improves the ontology relationship network for oil and gas exploration and development (Table 11).

**Table 11.** Example of the ontology relationship network for oil and gas exploration and development.

| Relationship Name | Description | Example |
|---|---|---|
| Belong to | A belongs to B, or B contains A | The application area belongs to a region |
| Be in | The position of A is B | Well No. 1 is located at the high part of the buried hill anticline |
| Drilling encounter | Formation encountered by bit A | Well No. 1 Drilling Encounters Quaternary System |
| Sub event | A is a sub event of B | Drilling “sub event” drilling |
| Nearby | A and B can be discovered and found together | Well No. 1 near Well No. 2 |

### 4.2 Exploration and Development Knowledge Graph Ontology Model

By defining ontology concepts and constructing an exploration and development sector domain ontology, 258 ontology concepts and 244 relationships for oil and gas exploration and development sector services were completed (Fig. 8).

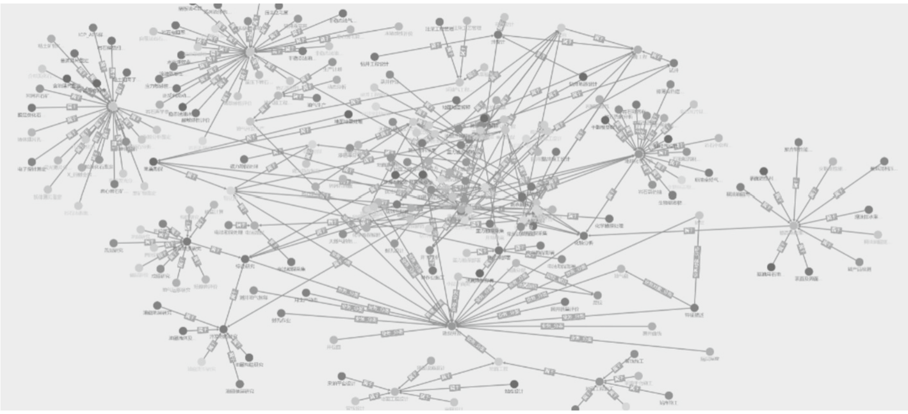


**Fig. 8.** Example of ontology model.

## 5 Conclusion

The construction of domain knowledge graph ontology of exploration and development is a process of continuous development. With the continuous updating of new technologies, methods and data of exploration and development, the construction of knowledge graph is also a process of continuous updating. Ontology needs to be constantly improved and developed to ensure the accuracy and integrity of exploration and development knowledge graph ontology. In addition, although this article describes the construction methods and ideas of the exploration and development knowledge graph, there is currently no unified and standardized process for constructing the ontology in the exploration and development field, and continuous exploration and research are needed to promote the continuous development of the oil and gas exploration and development sector knowledge graph.

Aiming at the problem of domain knowledge expression and application of exploration and development sector, this paper analyzes the knowledge graph technology from three aspects of sector domain, object domain and feature domain, proposes the construction method and process of domain knowledge graph ontology of exploration and development, and solves the expression problem of domain knowledge of exploration and development. Based on the knowledge graph and artificial intelligence platform technical specifications, referring to the EPDM model, design a construction plan for the exploration and development knowledge graph, and ultimately form an exploration and development ontology model.

(1) Based on the core sector knowledge of oil and gas exploration and development, knowledge related to exploration and development sector is automatically extracted from professional websites, books, and other materials, and a professional vocabulary and feature corpus in the field of oil and gas exploration and development are established. The professional vocabulary and corpus should cover all professional vocabulary or knowledge points in the exploration and development field as much as possible, without emphasizing the logical relationship between knowledge points, and emphasizing whether to fully cover all important concepts in the exploration and development field, which is a standardized collection of concepts for establishing an exploration and development knowledge graph.

(2) Taking sector, objects and characteristics as the main line, based on the knowledge graph of exploration and development and the technical specification of artificial intelligence platform, referring to the EPDM model, and following the principles of sector integrity, clear sector structure, unique sector classification, sector description specification and good scalability, a classification system of exploration and development domain knowledge is established. Through scientific, complete and clear expression of the domain knowledge of exploration and development and the relationship between knowledge, a knowledge classification covering the exploration and development domain is formed.

(3) On the basis of the constructed knowledge classification system, based on the exploration and development knowledge graph and the technical specifications of the artificial intelligence platform, the ontology construction language is used to express knowledge clearly and clearly, and the exploration and development domain ontology is established from the three dimensions of sector, object and feature, completing the preliminary transformation of the exploration and development knowledge classification system to the knowledge graph ontology.

**Acknowledgments.** The work is supported by the following projects: ①CNPC Project “Research on Intelligent Interpretation Technology and Knowledge Graph of Oil and Gas Exploration” (Number: 2021DJ7003); ②CNPC RIPED project “Research on integrated mathematical model of fracture network and software module development” (Number: YGJ2019–07-04). ③CNPC major scientific and technological project of upstream key core technology "Volume fracturing optimization design software" sub-project "EDFM-based Component reservoir simulation software" (Number: 2020B-4118). ④CNPC RIPED project “Joining 5 industrial alliance organizations such as the University of Calgary for exchange and cooperation research” sub-project “Joining the MMRD Consortium at Colorado School of Mines” (Number: 2021DQ0105–03). And in the process of completing this paper, the author has received technical support and help from researchers, including Professor Ren-bin Gong of Technology Consult Center of Research Institute of Petroleum Exploration & Development of PetroChina, Xiao-guo Wan and Bao-zhu Guan of China Petroleum Tarim Oil and Gas Field Branch Exploration and Development Research Institute, and experts, incluidng Tao Su, Ji-qing Wang, Xing-yin Zhang of Huawei Technology Co., Ltd, and the author would like to express the sincere gratitude.

## References

1. Liu, Q., Li, Y., Yang, D.H., et al.: Overview of knowledge graph construction technology. Meter Comput. Res. Dev. **53**(3), 582–600 (2016)
2. Xu, Z., Sheng, Y., He, L., Wang, Y., et al.: Overview of knowledge graph technology. J. Univ. Electron. Sci. Technol. **45**(4), 589–606 (2016)
3. Wen, B., Li, Z.: Research on the construction method of ontology in the field of petroleum exploration and development. Comput. Eng. Appl. **45**(34), 1–4 (2009)
4. Zhou, S., Yan, J., Bao, H., et al.: Ontology construction and application in the field of petroleum exploration and development. Comput. Syst. Appl. **24**(5), 172–176 (2015)
5. Du, R., Shang, F., Wu, Y.: Research on ontology based construction of domain knowledge in petroleum development. Sci. Technol. Eng. **10**(19), 4656–4662 (2010)
6. Wang, C.: Research on the Construction and Evolution Method of Ontology Knowledge Base. Ocean University of China, Qingdao (2008)
7. Yuan, G., Chen, S., Xin, Y., Deng, X.: Research on the application of ontology construction theory in the petroleum field. Comput. Technol. Autom. **03**, 113–118 (2011)
8. Qi, H., Dong, S., Zhang, L., Hu, H., Fan, J.: Construction and prospects of earth science knowledge atlas. J. Univ. Geol. **26**(1), 002–010 (2020)
9. Lin, J., Li, X.-W.: Classification and thinking of oil and gas reservoir exploration stages. Daqing Petrol. Geol. Dev. **36**(04), 20–25 (2017)
10. Nian, Z., Liang, S., Ma, F., Li, S.: Research and application of knowledge representation methods. Comput. Appl. Res. (05), 234–236+286 (2007)
11. Chen, H.: Research on Ontology Based Knowledge Representation. Changsha University of Technology, Changsha (2006)
12. Zhang, X.: Analysis and grabbing implementation of web crawler based on scrapy framework. Comput. Program. Skills Maint. (02), 18–19, 44 (2022)
13. Li, N.-F.: Research on Semantic Association Mechanism Based on Petroleum Domain Ontology. Northeast Petroleum University, Daqing (2014)
14. Wang, X., Deng, D., Meng, X., Tang, X., Guo, P., Lin, C.S.: Knowledge acquisition model and implementation for oil and gas exploration and development based on domain ontology. J. Northeast Petrol. Univ. **40**(04), 74–79, 87–89 (2016)
15. Zhang, X., Ji, Z., Nian, L., Xiao, L., Jing, N.: Preliminary study on the construction of petrochemical domain knowledge ontology. Inf. Syst. Eng. **03**, 27–28 (2012)
16. Duan, L.: Research on Automatic Ontology Construction Meth-od in the Petroleum Field Based on Text Analysis. Northeast Petroleum University, Beijing (2015)
17. Li, F., Mao, Z.: Research on application ontology construction method and case analysis. Libr. Sci. Res. (19), 31–41 (2014)
18. Gong, R., Li, X., Li, N., et al.: Artificial Intelligence for Oil and Gas, pp. 9–10. Petroleum Industry Press, Beijing (2021)
19. Li, N., Gong, R.B., Li, X., et al.: Factor analysis of affecting the accuracy for intelligent picking of seismic first arrivals with deep learning model. In: Lin, J. (ed.) Proceedings of the International Field Exploration and Development Conference 2022. IFEDC 2022. Springer Series in Geomechanics and Geoengineering. Springer, Singapore (2022). https://doi.org/10.1007/978-981-99-1964-2_598
20. Li, N., Gong, R., Liu, Z., et al.: Application of artificial intelligence technology in single well production and water cut prediction. In: Lin, J. (ed.) Proceedings of the International Field Exploration and Development Conference 2021. IFEDC 2021. Springer Series in Geomechanics and Geoengineering. Springer, Singapore (2021). https://doi.org/10.1007/978-981-19-2149-0_47

# A Method for Automatic Identification of Natural Fracture Based on Machine Learning: A Case Study on the Dahebian Block of the Liupanshui Basin in Guizhou Province

Wei-guang Zhao[1,4], Shu-xun Sang[1,2,3](✉), De-qiang Cheng[5], Si-jie Han[2,3],
Xiao-zhi Zhou[1,6], Jin-chao Zhang[1], and Fu-ping Zhao[7,8]

[1] School of Resources and Geosciences, China University of Mining and Technology, Xuzhou 221116, China
shxsang@cumt.edu.cn

[2] Jiangsu Key Laboratory of Coal-Based Greenhouse Gas Controland Utilization, China University of Mining and Technology, Xuzhou 221008, China

[3] Carbon Neutrality Institute, China University of Mining and Technology, Xuzhou 221008, China

[4] Artificial Intelligence Research Institute, China University of Mining and Technology, Xuzhou 221116, China

[5] School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China

[6] Key Laboratory of Coalbed Methane Resources and Reservoir Formation Process, Ministry of Education, China University of Mining and Technology, Xuzhou 221116, China

[7] Key Laboratory of Unconventional Natural Gas Evaluation and Development in Complex Tectonic Areas, Ministry of Natural Resources, Guiyang 550009, China

[8] Guizhou Engineering Research Institute of Oil & Gas Exploration and Development, Guiyang 550009, China

**Abstract.** Natural fractures are effective storage spaces and important seepage channels for oil and gas reservoirs. Accurately identifying natural fractures in reservoirs is crucial for the exploration and development of oil and gas resources. This article combines conventional and imaging logging data and uses machine

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20–22 September 2023.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 461–472, 2024.
https://doi.org/10.1007/978-981-97-0272-5_39

learning to automatically identify natural fractures in reservoirs. The fracture labels of conventional logging come from imaging logging. Conventional logging data is decomposed through multi-scale wavelet to extract components that reflect fracture information, and further build the original data set. The AdaBoost model is trained based on a modified dataset of balanced samples for automatic fractures recognition in logging. The research results indicate that the approximate component and high-frequency component reflect the fluctuation of the formation and noise information respectively, and have little impact on the reservoir fractures identification; The medium frequency component can reflect the characteristic information of fractures and can be used for model training; After hyper-parameter optimization, the AdaBoost model has high accuracy and generalization ability, and can still accurately identify the types and distribution of natural fractures from the actual unbalanced logging data. This research has important guiding significance for the accurate characterization and construction of reservoir.

**Keywords:** Conventional Logging · Imaging Logging · Natural Fractures Identification · Machine Learning · Wavelet Transform

## 1 Introduction

Natural fracture is an important factor affecting reservoir characteristics. The distribution and development of fractures are closely related to the stable and high production of oil and gas resources [1, 2]. Therefore, fracture identification and evaluation is the basic research content for efficient exploration of oil and gas resources [3]. The accurate identification of natural fractures is helpful for people to understand the physical characteristics of the reservoir, to select reasonable mining technology or reservoir reconstruction scheme [4, 6].

Geophysical logging has been widely used in natural fracture identification because of its fine vertical resolution. It mainly includes conventional logging and special logging. Fractures produce special responses on conventional logging curves, so it is feasible to identify fractures based on this logging [7]. Aghli et al. [8]used the first derivative curve of conventional logging curve to detect the development of reservoir fractures, and the results had a good correspondence with the fractures revealed by imaging logging. Tokhmchi et al. [9] found a good correspondence between the energy and fracture density of conventional logging curves, and established a nonlinear relationship between them. Imaging logging is a special kind of logging. Compared with conventional logging, it can more intuitively show the shape and distribution of fractures [10], can control the scaling and has been widely used in fracture identification and evaluation [11, 12]. However, the high cost of imaging logging makes it unable to be widely used. Therefore, conventional logging is still the main data source for fracture identification.

Because of the complex structure and serious heterogeneity of the reservoir, it is difficult to determine the direct relationship between fractures and conventional logging. Machine learning can deeply mine the fracture information in conventional logging data and establish a reasonable mapping relationship [13]. In recent years, classical machine learning models such as multi-layer perceptron and random forest have been applied to the identification and evaluation of fractures [14, 15]. Conventional logging data is relatively easy to obtain, but it cannot directly reflect the actual distribution of fractures, and cannot provide accurate labels for data samples. It is a feasible scheme to combine conventional logging with imaging logging for accurate identification of natural fractures [16, 17]. Therefore, this paper proposes a method of natural fracture identification using AdaBoost model based on conventional logging and imaging logging data.

This study mainly identifies open fractures and fill fractures. The fractures identified by imaging logging are used to mark the samples of conventional logging. Then the conventional logging data is decomposed by wavelet. The wavelet components associated with fractures are used to construct data sets for AdaBoost model training. The trained model is used to identify natural fractures. This sturdy provides an efficient and convenient method for the reservoir natural fractures identification, which has important engineering application value.

## 2 Geological Setting

The study area is located in the Dahebian syncline of Liupanshui City, Guizhou Province, belonging to the Yangtze Landmass (grade I), Qianbei uplift (grade II), Liupanshui fault depression (grade III) and Weining northwest structural deformation area. The strata include the Emeishan basalt formation of Upper Permian ($P_3\beta$), Longtan Formation($P_3l$),Lower Triassic Feixianguan Formation($T_1f$), Yongningzhen Formation($T_1yn$),Middle Triassic Guanling Formation($T_2g$)and quaternary system(Q)overlying the above strata. The main lithology includes siltstone, argillaceous siltstone, mudstone, coal and basalt. Fractures are developed in various rock layers, and some fractures are fill with calcite and other minerals (see Fig. 1).

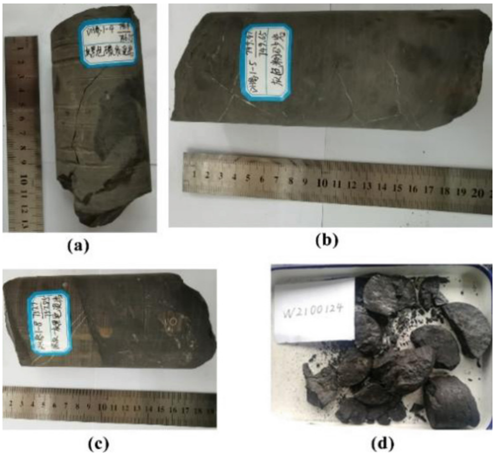

(a) (b) (c) (d)

**Fig. 1.** Core photos with fractures. (a) Mudstone with open fracture (b) siltstone with fill fractures (c) mudstone with open and fill fractures (d) coal

# 3 Methodology

## 3.1 Adaboost Algorithm

AdaBoost is an ensemble learning algorithm that promotes weak learners to strong learners [18]. The core principle is to fit a sequence of weak learners (such as decision trees). Adjust the weights of the training samples based on the performance of the weak learners generated during each training session, and apply them to the generation of the next learner until all weak learners have been trained. Then, the predictions of all weak learners are combined through weighted majority vote (or summation) to produce the final prediction. During the iteration process, the weights of training samples that were incorrectly predicted will increase, the weights of training samples that were correctly predicted will decrease. In this way, each subsequent weak learner focuses on the example which the previous learner identifying errors. The iterative approach adopted by the Adaboost algorithm enhances the accuracy and generalization ability of the classifier [19].

## 3.2 Wavelet Transform

Wavelet transform is a time-frequency localization analysis method with a variable time-frequency window [20].The formula is given by Eq. (1)

$$WT(\alpha, \tau) = \frac{1}{\sqrt{\alpha}} \int_{-\infty}^{\infty} f(t)\varphi \cdot \left(\frac{t-\tau}{\alpha}\right)dt \tag{1}$$

where, $f(t)$ is the original signal, $\varphi$ is a base wavelet transform function, $\alpha$ is the scale factor, $\tau$ is the shift factor [21]. $\alpha$ and $\tau$ can control the scaling and translation of wavelets. The original signal can be decomposed into approximate components and multi-level detail components by wavelet transform. The approximate component represents the low-frequency information of the signal, and the detail components represent the information of different frequencies. Fractures can disturb logging signals, but this disturbance may be masked by the overall formation signal. At the same time, the noise will also affect the disturbance. Through the multi-scale wavelet decomposition of the logging signal, the signal characteristics of different frequencies can be obtained, so as to extract the critical component that can best reflect the fracture.This method has been successfully applied in lithology division and fracture identification [22, 23].

## 3.3 Adaptive Synthetic (ADASYN) Algorithm

In this study, the natural fractures to be identified include open fractures and fill fractures. The samples in the original data set are from conventional logging, and the fracture labels of the samples are marked by imaging logging. Non-fracture samples are marked as 0, open fracture samples are marked as 1, and fill fracture samples are marked as 2. The dataset contains a total of 6641 samples, including 6376 non fracture samples, 114 open fracture samples, and 151 fill fracture samples. In the original data set, the number of samples marked as fractures accounts for a low proportion of the total samples. If the learner determines all samples as non-fracture samples, the overall accuracy of the model is still as high as 96%. Imbalanced data will cause the model to pay more attention to the learning of majority classes and cannot learn the characteristics of minority classes. A large number of fracture samples will be judged as non-fracture samples, so the fractures cannot be identified. The omission and misjudgment of fractures have a significant impact on the accurate characterization of reservoir structure, and then affect the subsequent decision-making. An ideal learner should recognize the fracture samples as accurately and comprehensively as possible. Therefore, the original data set must be modified to solve the problem of data imbalance, so as to improve the recognition ability of the learner for fractures. ADASYN is an oversampling method used to deal with unbalanced data sets. The algorithm is similar to (Synthetic Minority Oversampling Technique) SMOTE [24], but it generates different number of samples depending on an estimate of the local distribution of the class to be oversampled [25]. Through ADASYN algorithm, new minority samples are generated to achieve sample equilibrium. There are 19111 samples in the new data set after reconstruction, including 6376 non-fracture samples, 6370 open fracture samples and 6365 fill fracture samples.

### 3.4 Tree Structured Parzen Estimator

Tree structured Parzen Estimator (TPE) is a Bayesian optimization modeling strategy aimed at finding the optimal parameters from the configuration space to maximize the expected improvements (EI) [26]. TPE algorithm can accurately and quickly find the optimal configuration parameters of AdaBoost model.

### 3.5 Model Training and Performance Evaluation

The workflow of model training is shown in Fig. 2. The data set was divided into training set (75%) and testing set (25%). The training set is used for model training and hyper-parameter optimization, and the testing set is used for model performance evaluation. The parameters to be optimized mainly include: n_estimators,learning_rate, max_depth and min_samples_split. Where, max_ depth and min_ samples_split is the configuration parameter of the base learner (decision tree). To prevent over fitting during training, the average error loss of 5-fold cross-validation is used as the objective function of the hyper-parameter optimization. After obtaining the optimal parameter configuration through iterative training, the performance of AdaBoost model is evaluated based on the testing set.

## 4 Results and Discussion

### 4.1 Fracture Feature Extraction by Wavelet Transform

In this study, the original logging data included the natural gamma (GR), sonic (AC), compensated neutron (CNL), densities (DEN), deep lateral resistance (LLD), caliper logging (CAL), spontaneous potential(SP).In order to eliminate the dimensional difference between different logs and improve the speed of machine learning, the data are

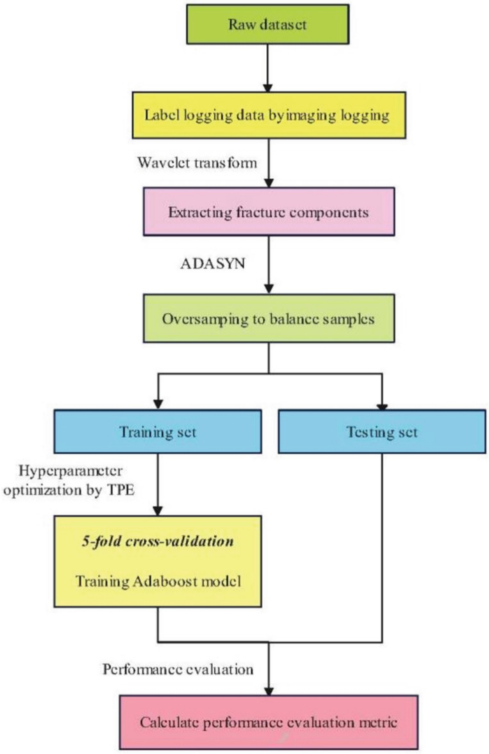

Raw dataset
Label logging data byimaging logging
Wavelet transform
Extracting fracture components
ADASYN
Oversamping to balance samples
Training set
Testing set
Hyperparameter optimization by TPE
5-fold cross-validation
Training Adaboost model
Performance evaluation
Calculate performance evaluation metric

**Fig. 2.** The workflow of model training

standardized. Next, all logging data are decomposed by 5-level multiscale wavelet based on db8 wavelet function. Taking GR as an example, the approximate component and the each level detail component are shown in Fig. 3.

The cA5 is an approximate component of GR curve, representing low-frequency information, reflecting the overall fluctuation of the formation, and is less sensitive to fracture identification. The cD5, cD4, cD3, cD2, cD1 are the detail components of GR curve. cD5 and cD4,are medium frequency components, which have obvious response to fractures.cD3、cD2 and cD1 are high-frequency components, representing noise or other interference unrelated to fractures. Through the multi-scale wavelet transform of logging curves, it can be considered that cD5 and cD4 record the characteristic information of fractures, so the sum of these two components is used to establish the data set of machine learning [22].

### 4.2 Hyper-parameter Optimization

The data set established by wavelet component is modified by ADASYN algorithm to achieve sample equalization. The new dataset contains 19111 samples. It is used for AdaBoost model training, optimization and evaluation. 75% of the data is used as training set for model training and hyper-parameter optimization. After 1000 iterations

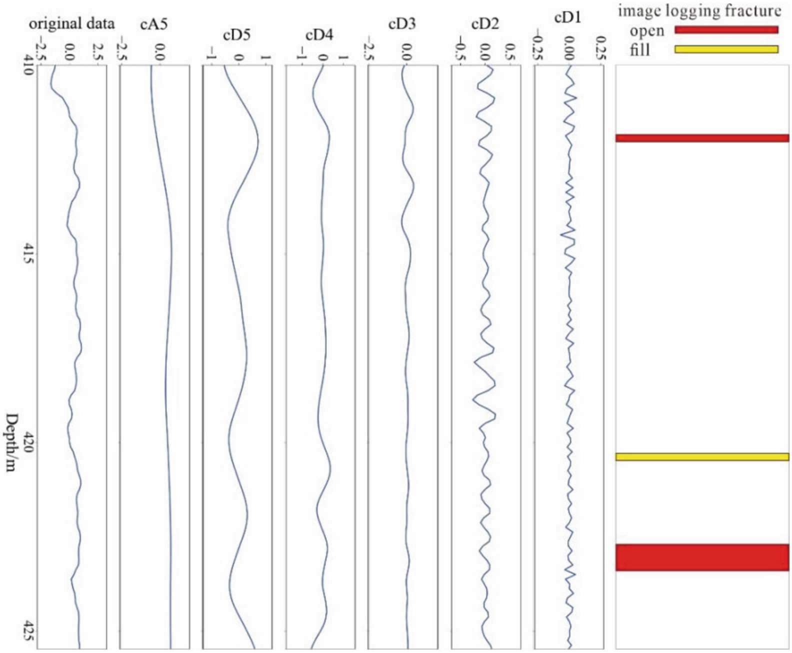


**Fig. 3.** GR wavelet transform

of TPE algorithm, the optimal hyper parameter configuration of the model is finally determined (Table 1).

**Table 1.** Optimal configuration of the hyper-Parameters

| Hyper-parameters | Description | Range | Optimal value |
|---|---|---|---|
| n_estimators | The maximum number of estimators at which boosting is terminated | (50,200) | **118** |
| learning_rate | Weight applied to each classifier at each boosting iteration | (0,1) | **0.572** |
| max_depth | Base estimator parameter. The maximum depth of the tree | (1,6) | **6** |
| min_samples_split | Base estimator parameter. The minimum number of samples required to split an internal node | (0,1) | **0.002** |

### 4.3 Performance Evaluation Result

The remainder 25% of the data is used to evaluate the performance of the model. In this paper, the overall accuracy (OA), precision, recall and F1-score. The calculation formula of other metric is as follows:

$$\mathrm{OA} = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

$$\mathrm{Precision} = \frac{TP}{FP + TP} \tag{3}$$

$$\mathrm{Recall} = \frac{TP}{FN + TP} \tag{4}$$

$$\mathrm{F1} = \frac{2{*}Recall{*}Precision}{Recall + Precision} \tag{5}$$

where, TP, TN, FN and FP are the four elements in the confusion matrix:true positive, true negative, false negative, false positive

Table 2 shows the calculation results of various evaluation metric of AdaBoost model with optimal hyper-parameter configuration on the testing set. After parameter optimization, the evaluation metrics of the model were improved from less than 0.65 to 0.98. The accuracy and generalization ability of the model have been significantly improved. At the same time, this study also trained Support Vector Machine (SVM), Multi layer Perceptron (MLP), and Random Forest (RF) based on the same training set, and evaluated them in the same test set. The evaluation results show that the evaluation metrics of the optimized AdaBoost model are higher than those of other models. Therefore, it can be proved that AdaBoost model can more accurately identify fractures.

**Table 2.** Performance measure of models

| Model | Measurement evaluation metric | | | |
|---|---|---|---|---|
| | OA | Precision | Recall | F1-score |
| SVM | 0.774 | 0.793 | 0.777 | 0.768 |
| MLP | 0.826 | 0.828 | 0.828 | 0.824 |
| RF | 0.813 | 0.830 | 0.815 | 0.813 |
| AdaBoost(default) | 0.642 | 0.641 | 0.644 | 0.641 |
| **AdaBoost(optimal)** | **0.982** | **0.983** | **0.983** | **0.983** |

### 4.4 Natural Fracture Identification

In this paper, in order to solve the problem of sample imbalance, the original data set is modified by ADASYN algorithm. The number, characteristics and distribution of samples in the new dataset are different from the original dataset. In order to test whether the optimized AdaBoost model can still accurately identify fractures from unbalanced data sets, the original unbalanced samples are put into the model to identify natural fractures. The identification results of natural fractures are shown in Fig. 4.

Figure 4 shows that AdaBoost model can still accurately identify the category, location and distribution of natural fractures in the original unbalanced data set. Figure 5 shows the confusion matrix of AdaBoost recognition results in the original data set. In the imbalance data set, although the proportion of fracture samples is only 4%, the model can still accurately identify most fractures. Therefore, the results can prove that AdaBoost model can directly use the logging components after wavelet decomposition to accurately identify natural fractures.

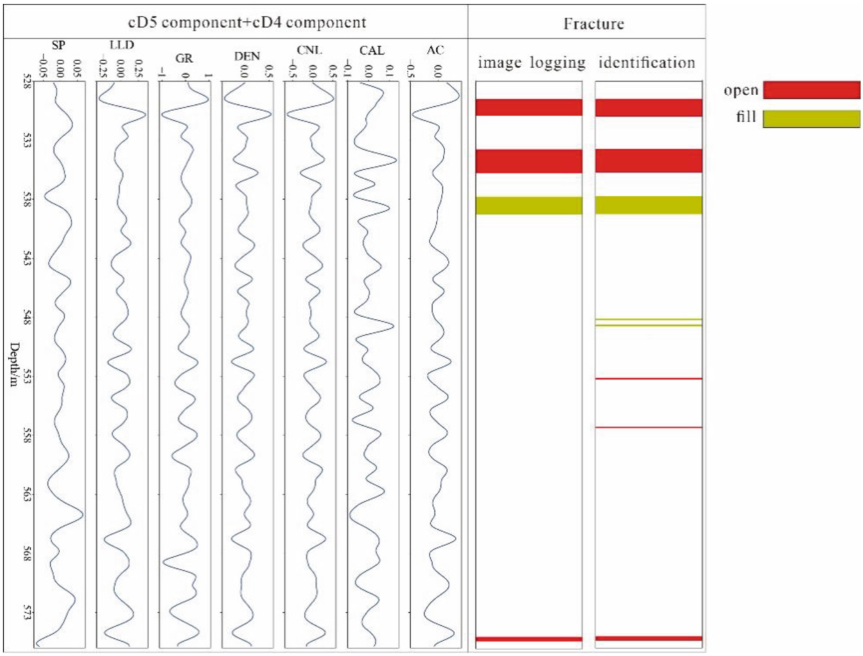


**Fig. 4.** Natural fracture identification results

In previous studies, drilling core is the most effective evidence to identify fractures. However, It cannot be widely used due to the limitation of cost and recovery factor [27], and cannot provide detailed fracture labels for samplesIn this study, the sample characteristics of the dataset used to train the AdaBoost model come from conventional logging. Imaging logging is used to mark fractures in conventional logging samples. The advantages of this processing method are as follows: 1) imaging logging can provide detailed natural fracture information, which solves the problem of fracture marking in conventional logging. 2) It is only necessary to extract cD5 + cD4 wavelet components from conventional logging data and put them into AdaBoost model. The type, location and distribution of natural fractures in this logging can be identified. Therefore, this study provides an efficient, practical and accurate method for the characterization and construction of reservoirs, and has important theoretical and application significance for the exploration and development of oil and gas resources and reservoir reconstruction.

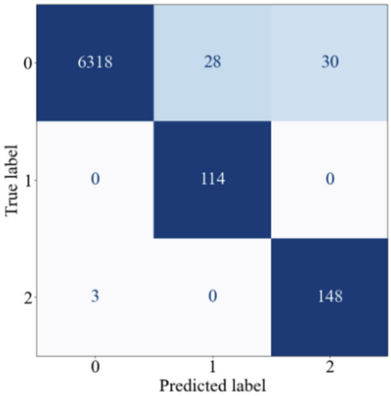


**Fig. 5.** Confusion matrix. 0: non-fracture, 1: open fracture, 2: fill fracture

# 5 Conclusions

In this study, AdaBoost model is constructed based on conventional logging and imaging logging, and natural fractures are identified. The main conclusions are as follows:

(1) The wavelet transform results of the original log show that cD5 and cD4 components have a greater response to fractures. The cD5 and cD4 can be considered to record the characteristic information of fractures. Other components represent the overall stratum fluctuation or noise, and are not related to fractures.

(2) The performance evaluation result of AdaBoost model with default parameter configuration on the test set is low. After hyper-parameter optimization, the performance of model has been greatly improved, and the accuracy and generalization ability have been significantly promoted.

(3) Although the AdaBoost model is trained based on the modified data set, the model can still directly identify fracture (minority class samples) from the imbalanced data set. Therefore, it proves that the model can accurately identify the type, location and distribution of natural fractures based on conventional logging data after wavelet transform.

**Acknowledgments.** The project is supported by the National Natural Science Foundation of China (Grant No. 42030810).

# References

1. Prasun, S., Wojtanowicz, A.K.: Semi-analytical prediction of critical oil rate in naturally fractured reservoirs with water coning. J. Petrol. Sci. Eng. **180**, 779–792 (2019)
2. Nie, X., Zou, C., Pan, L., et al.: Fracture analysis and determination of in-situ stress direction from resistivity and acoustic image logs and core data in the Wenchuan earthquake fault scientific drilling borehole-2 (50–1370m). Tectonophysics **593**, 161–171 (2013)

3. Shafiei, A., Dusseault, M.B., Kosari, E., et al.: Natural fractures characterization and in situ stresses inference in a carbonate reservoir-an integrated approach. Energies **11**(2), 312 (2018)
4. Chen, S.Y., Wang, Y.J., Guo, J.Y., et al.: Multi-scale evaluation of fractured carbonate reservoir and its implication to sweet-spot optimization: a case study of Tazhong oilfield, central tarim basin. China. Energy Reports **7**, 2976–2988 (2021)
5. Pan, D.D., Li, Y.H., Lin, C.J., et al.: Intelligent rock fracture identification based on image semantic segmentation: methodology and application. Environ. Earth Sci. **82**(3), 71 (2023)
6. Lai, J., Wang, G., Pang, X., et al.: Effect of pore structure on reservoir quality and oiliness in eocene dongying formation sandstones in Nanpu Sag, Bohai Bay Basin. Eastern China. Energy Fuel **32**(9), 220–9232 (2018)
7. Lyu, W.Y., Zeng, L.B., Liu, Z.Q., et al.: Fracture responses of conventional logs in tight-oil sandstones: a case study of the Upper Triassic Yanchang formation in southwest Ordos Basin. China. AAPG Bull **100**(9), 1399–2141 (2016)
8. Aghli, G., Soleimani, B., Moussavi-Harami, R., et al.: Fractured zones detection using conventional petrophysical logs by differentiation method and its correlation with image logs. J. Petrol. Sci. Eng. **142**(8), 152–162 (2016)
9. Tokhmchi, B., Memarian, H., Rezaee, M.R.: Estimation of the fracture density in fractured zones using petrophysical logs. J. Petrol. Sci. Eng. **72**(1), 206–213 (2010)
10. Rajabi, M., Sherkati, S., Bohloli, B., et al.: Subsurface fracture analysis and determination of in-situ stress direction using FMI logs: an example from the Santonian carbonates (Ilam Formation)in the Abadan Plain. Iran. Tectonophysics **492**, 192–200 (2010)
11. Lai, J., Wang, G.W., Fan, Z.Y., et al.: Fracture detection in oil-based drilling mud using a combination of borehole image and sonic logs. Mar. Pet. Geol. **84**, 195–214 (2017)
12. Khoshbakht, F., Memarian, H., Mohammadnia, M.: Comparison of Asmari, Pabdeh and Gurpi formation's fractures, derived from image log. J. Petrol. Sci. Eng. **67**(1–2), 65–74 (2009)
13. Dong, S.Q., Zeng, L.B., Lyu, W.Y., et al.: Fracture Identification by semi-supervised learning using conventional logs in tight sandstones of Ordos Basin, China. J. Nat. Gas Sci. Eng. **76**, 103–131 (2020)
14. Pei, J.Y., Zhang, Y.F.: Prediction of reservoir fracture parameters based on the multi-layer perceptron machine-learning method: a case study of Ordovician and Cambrian carbonate rocks in Nanpu Sag, Bohai Bay Basin. China. Process. **10**(11), 2445 (2022)
15. Bhattacharya, S., Mishra, S.: Applications of machine learning for facies and fracture prediction using Bayesian network theory and random forest: case studies from the Appalachian basin, USA. J. Petrol. Sci. Eng. **170**, 1005–1017 (2018)
16. Amir, M., Ali, K., David, A.W., et al.: Natural fractures characterization by integration of FMI logs, well logs and core data: a case study from the Sarvak Formation (Iran). J. Petrol. Explor. Prod. Technol. **13**, 1247–1263 (2023)
17. Qiu, X.L., Tan, C.Q., Lu, Y.Y., et al.: Evaluation of fractures using conventional and FMI logs, and 3D seismic interpretation in continental tight sandstone reservoir. Open Geosci. **14**, 530–543 (2022)
18. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55(1), 119–139 (1997)
19. Kevin, W.W., Jiang, Z.H.: Application of adaptive boosting (AdaBoost) in demand-driven acquisition(DDA) prediction: a machine-learning approach. J. Acad. Librariansh. **45**(3), 203–212 (2019)
20. Perez-Mun˜oz, T., Velasco-Hernandez, J., Hernandez-Marti-nez, E.: Wavelet transform analysis for lithological characteristics identification in siliciclastic oil fields. J. Appl. Geophys. **98**, 298–308 (2013)
21. Goupillaud, R.A., Grossmann, A., Morlet, J.: Cycle-octave and related transform inseismic signal analysis. Geoexploration **23**, 85–102 (1985)

22. Chen, T.J., Ma,.G.D., Wang. X., et al.: Deformation degree estimate for coal seam using well logs as input: a case study. J. Environ. Eng. Geophys. **23**(1), 89–101 (2018)
23. Zhang, X.F., Pan, B.Z., Wang, F., Han, X.: A study of wavelet transforms applied for fracture identification and fracture density evaluation. Appl. Geophys. **8**(2), 164–169 (2011)
24. Chawla, N.V., Bowyer, K.W., Hall, L.O., et al.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
25. He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks, pp. 1322–1328 (2008)
26. Bergstra, J., Bardenet, R., Bengio, Y., et al.: Algorithms for hyper-parameter optimization. In: 25th International Conference on Neural Information Processing Systems (2011)
27. Aghli, G., Soleimani, B., Moussavi-Harami, R., et al.: Fractured zones detection using conventional petrophysical logs by differentiation method and its correlation with image logs. J. Petrol. Sci. Eng. **142**, 152–162 (2016)

# Development and Application of Wireless Strain Test System in the Bearing Capacity Test of Oil Derrick

Deng Jia[1,2](✉), Zhi-xiong Zhou[1], Xiao-guang Yang[1], Xue-li Luo[2], Yang Li[2], Ling Jin[2], Wei-dong Zuo[2], Na Zhang[1], and Ying Ma[1]

[1] CNPC Engineering Technology R&D Company Limited, Beijing, China
jiadeng123@qq.com
[2] Beijing Kembl Petroleum Technology Co. Ltd., Beijing, China

**Abstract.** In order to ensure the safe production of oil drilling equipment, about 2000 sets of in-service oil drilling rigs and workover rigs need to be tested and evaluated every year, covering Daqing, Changqing, southwest, Tarim, Dagang and other major oil and gas fields as the key equipment of the drilling rig system, the bearing capacity of the Derrick is directly related to the safety of production. In view of the complex data transmission line, low efficiency and high labor intensity of the wired strain tester, a wireless strain testing system with the function of testing and analyzing the bearing capacity of Derrick is developed the integration of wireless transmission of Derrick bearing capacity test data, on-line analysis and load capacity evaluation is realized. Through the quasi-calibration of Beijing Institute of Metrology and Field Test of Oil Derrick, it is proved that the research results have high accuracy in strain detection (JJG623–2005), strong stability, high efficiency, low power consumption, friendly interpersonal interface and broad application prospects.

**Keywords:** Oil Derrick · bearing capacity · Wireless Communication · strain Measurement

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 473–484, 2024.
https://doi.org/10.1007/978-981-97-0272-5_40

## 1 Introduction

Oil Derrick is a kind of tower mast steel structure, which is used for hoisting equipment and tools such as crane, traveling car, big hook, etc., to bear the load caused by drilling, casing and other operations, and to provide height and space for drilling and storing string operation [1–3]. Due to the poor working environment of the oil exploration and development site and the complex load on the Derrick, it is affected by various defects, overloading, corrosion and other factors in the process of long-term use, resulting in bad changes in the internal structure of the Derrick of the in-service drilling rig. as a result, the bearing capacity of the original design of the drilling rig Derrick is reduced [4–6].

The oil Derrick belongs to a large steel structure system. If the traditional wired strain measurement method is adopted, the wiring installation of the system is tedious, the workload is heavy, the measurement accuracy is affected by the environment and wire length, and the testing time is longer [7–9]. Compared with the traditional wired system, the wireless strain testing system is more convenient, more efficient, high reliability, low power consumption, high precision, strong stability and other advantages, so it is very suitable to be used in the field testing of oil Derrick.

## 2 Technical Principle

The wireless strain testing system is a set of precision instrument system for measuring the strain of metal materials or structures. The data is transmitted through the wireless network. In addition to the general speed measurement module, the software also integrates the analysis and evaluation module of the bearing capacity of the oil Derrick. The integration of data acquisition and analysis is realized. The instrument uses the wireless strain detection base station to complete the strain measurement with high precision and high stability; each detection base station can work independently and distribute; the wireless data transmission is realized through the wireless test network composed of the detection base station, the router and the computer; the computer is equipped with wireless strain testing system software to collect, store, display and analyze the received test data in real time to complete the strain test. The hardware system consists of several wireless strain detection base stations, a wireless WIFI router and a computer, and the software system is a wireless strain testing system. The networking structure of the system is shown in Fig. 1.

## 3 System Composition

The wireless strain test system is composed of 8 wireless strain test base stations, 8 sensor connecting cables, 1 wireless router and a set of wireless strain test software. The whole wireless strain test system is shown in Fig. 2

The overview diagram of the software interface is shown in Fig. 3. The whole program consists of menu bar, toolbar, base station navigation area, label function area and so on. The navigation area of the base station equipment is a tree view, which shows all the base stations and the corresponding test points currently online. When switching between different base stations, the display data of the corresponding tabbed function

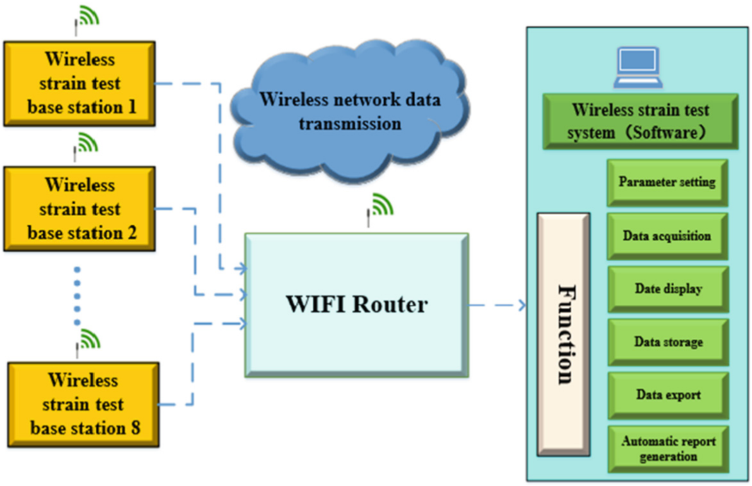

Wireless strain test base station 1
Wireless strain test base station 2
Wireless strain test base station 8
Wireless network data transmission
WIFI Router
Wireless strain test system（Software）
Function
Parameter setting
Data acquisition
Date display
Data storage
Data export
Automatic report generation

**Fig. 1.** The networking structure of the system

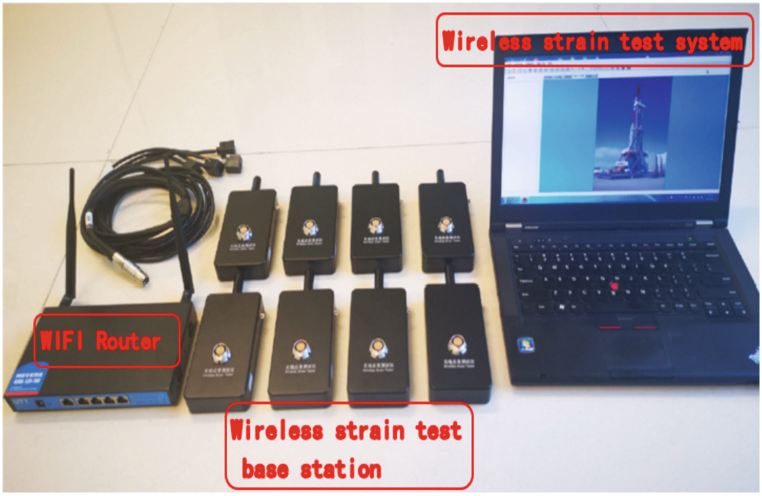
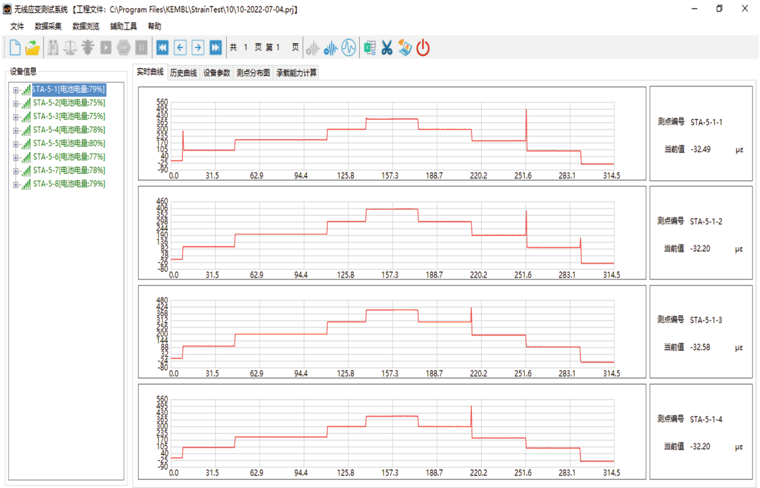

Wireless strain test system
WIFI Router
Wireless strain test base station

**Fig. 2.** Physical diagram of wireless strain testing system

area on the right will also be switched accordingly. The menu system contains all the functions of the system, including file menu module, data acquisition menu module, data browsing menu module, auxiliary tools menu module, help menu module and label function area module. Each wireless strain detection base station passes the distributed location test and completes the strain measurement independently. Common functions place shortcut buttons on the toolbar to improve operational efficiency.

By analyzing and comparing the structure composition, working principle and operating environment of drilling pump and winch equipment, the faults of drilling pump and winch are summarized as shown in Table 1 and Table 2.



无线应变测试系统 【工程文件：C:\Program Files\KEMBL\StrainTest\10\10-2022-07-04.prj】

设备信息

STA-5-1[电池电量:79%]
STA-5-2[电池电量:75%]
STA-5-3[电池电量:75%]
STA-5-4[电池电量:78%]
STA-5-5[电池电量:80%]
STA-5-6[电池电量:77%]
STA-5-7[电池电量:78%]
STA-5-8[电池电量:79%]

测点编号 STA-5-1-1
当前值 -32.49 με

测点编号 STA-5-1-2
当前值 -32.20 με

测点编号 STA-5-1-3
当前值 -32.58 με

测点编号 STA-5-1-4
当前值 -32.20 με

**Fig. 3.** Software interface overview diagram

## 4 Main Technical Index

The main technical performance indicators of wireless strain testing base stations are shown in Table 1.

The performance of the wireless strain testing system software is shown in Table 2.

The innovation of wireless strain testing system has the following two points.

1. The high-precision data acquisition and processing technology based on high-precision differential operational amplifier, low-noise temperature compensation, precision voltage stabilization and high-precision strain testing technology based on elastic elements in low temperature environment are integrated. A wireless static strain tester for on-line detection of drilling equipment is developed.
2. Combining the real-time data analysis model of bearing capacity with high-precision data acquisition system, a wireless static strain testing system with the function of real-time analysis of bearing capacity of drilling equipment is developed.
3. Strain testing and load capacity calculation are shown in Fig. 4。

**Table 1.** Technical index of wireless strain testing base station

| Technical parameters | Unit | Parameter value |
|---|---|---|
| Accuracy level | Stage | 0.1(JJG623–2005), ± (0.1%red ± 1 με) |
| Highest resolution | με | 1 |
| Zero drift | / | 0.1με/4h |
| Indication stability | / | 0.01%/4h |
| Sampling frequency | Hz | 2 |
| Bridge way | / | Full bridge, half bridge, 1/4 bridge optional |
| Data transmission form | / | Wifi transmission |
| Effective transmission distance | m | 100 |
| Number of channels | / | 4 channels/base station, supporting up to 32 base stations for synchronous acquisition (128 channels) |
| Working voltage | V | 4.2 |
| Working current | mA | 160 |
| Power supply mode | / | Lithium battery powered, continuous working time ≥30 h |
| Continuous storage time of test data | min | ≥20 |
| Strain range | με | ±5000 |
| Operating temperature | °C | −20~60 |
| Product size | mm | 1450 × 750 × 300 |
| Product weight | g | 350 |

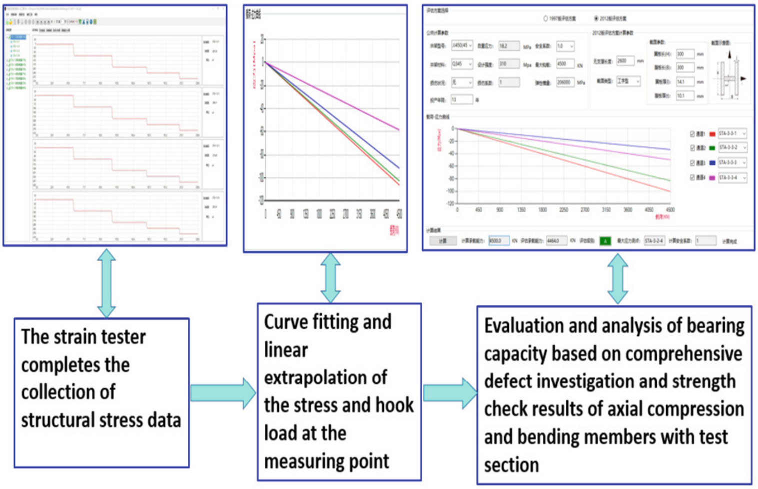

The strain tester completes the collection of structural stress data
Curve fitting and linear extrapolation of the stress and hook load at the measuring point
Evaluation and analysis of bearing capacity based on comprehensive defect investigation and strength check results of axial compression and bending members with test section

**Fig. 4.** Combination of strain test and bearing capacity evaluation

**Table 2.** Technical index of wireless strain testing system software

| Software name | Wireless strain testing system |
|---|---|
| Function | 1. Single measurement data processing and data storage upper limit 100 M;<br>2. Excel format output of arbitrary sampling data can be completed;<br>3. The software can automatically generate and export curves;<br>4. Data is transmitted over wireless network and TCP/IP protocol to ensure stable and reliable data transmission;<br>5. The collected data are displayed graphically, intuitively and clearly;<br>6. Complete the calculation of bearing capacity by collecting data;<br>7. One-click configuration of wireless network card parameters |
| Development environment | The development environment of the system is Windows 10, the development tool is Visual Studio, the programming language is Clover, and the running environment is 64-bit Windows 7, Windows 8 and Windows 10 systems |
| Operation environment | CPU: above 1 GHz<br>Memory: more than 1 GB<br>Remaining disk space: more than 10 GB<br>Graphics card: standard 16-bit VGA or above<br>Minimum screen resolution is 1366 × 768<br>100 M wireless Ethernet |

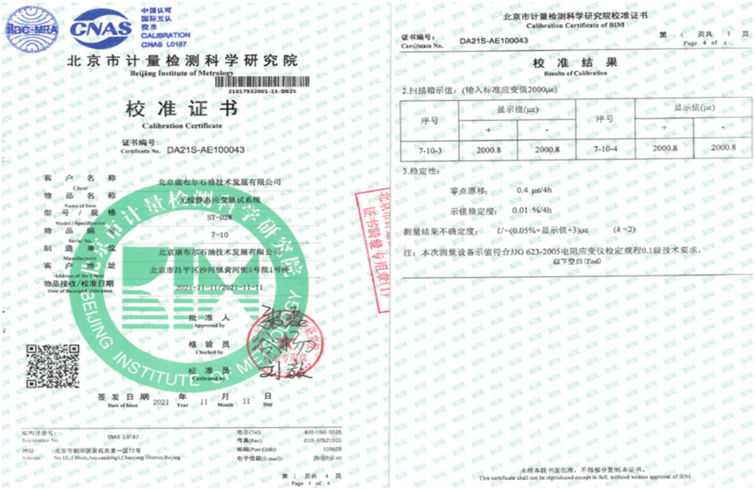
# 5 Test

## 5.1 Laboratory Test

The wireless strain testing system is brought to Beijing Institute of Metrology and testing for testing and evaluation, and the tester selects the DR8 standard analog strain calibration instrument in the laboratory to compare the strain test data [10–12]. According to the verification regulation of JJG623–2005 resistance strain gauge, the indication error, zero drift and indication stability are verified and calibrated. The host is shown in Table 3. After the verification of the Institute of Metrology and the calibration certificate, the accuracy level of this wireless strain testing system meets the requirements of the verification regulation of JJG623–2005 resistance strain gauge, the indication error is (0.1%red + 1$\mu\varepsilon$), and the indication stability is 0.01%/4h. The Calibration-certificate of Beijing Institute of metrology is shown in Fig. 5.

**Table3.** Wireless strain testing base station (7–10) indication value

| Actual value/με | Test value /με | | Actual value /με | Test value /με | |
|---|---|---|---|---|---|
| | + | - | | + | - |
| 10 | 10.1 | 10.1 | 400 | 400.1 | 400.1 |
| 20 | 20.1 | 20.1 | 500 | 500.2 | 500.2 |
| 30 | 30.1 | 30.1 | 600 | 600.5 | 600.5 |
| 40 | 40.1 | 40.1 | 700 | 700.1 | 700.1 |
| 50 | 50.2 | 40.2 | 800 | 800.5 | 800.5 |
| 60 | 60.3 | 60.3 | 900 | 900.4 | 900.4 |
| 70 | 70.1 | 70.1 | 1000 | 1000.6 | 1000.6 |
| 80 | 80.0 | 80.0 | 2000 | 2000.8 | 2000.8 |
| 90 | 90.2 | 90.2 | 3000 | 3000.9 | 3000.9 |
| 100 | 100.3 | 100.3 | 4000 | 4001.2 | 4001.2 |
| 200 | 200.2 | 200.2 | 5000 | 5001.3 | 5001.3 |


中国认可
国际互认
校准
CALIBRATION
CNAS L0187

北京市计量检测科学研究院
Beijing Institute of Metrology

21017932001-11-D825

校准证书
Calibration Certificate

证书编号:
Certificate No. DA21S-AE100043

| | |
|---|---|
| 客户名称<br>Client | 北京康布尔石油技术发展有限公司 |
| 物品名称<br>Name of Item | 无线静态应变测试系统 |
| 型号/规格<br>Model / Specification | ST-02W |
| 物品编号<br>Serial No. | 7-10 |
| 制造单位<br>Manufacturer | 北京康布尔石油技术发展有限公司 |
| 客户地址<br>Address of the Client | 北京市昌平区沙河镇黄河街5号院1号楼 |
| 物品接收/校准日期<br>Date of Receipt/Calibration | 2021-11-11/2021-11-11 |

批准人
Approved by

核验员
Checked by

校准员
Calibrated by

签发日期 2021 年 11 月 11 日
Date of Issue Year Month Day

北京市计量检测科学研究院
BEIJING INSTITUTE OF ME

证书骑缝专用章(1)

| | | | |
|---|---|---|---|
| 机构注册号:<br>Registration No. | CNAS L0187 | 电话(Tel): | 400-090-0026 |
| | | 传真(Fax): | 010-58821500 |
| 地址:<br>Address: | 北京市朝阳区安苑东里一区12号<br>No.12, 1 block, Anyuandongli, Chaoyang District, Beijing | 邮编(Post Code): | 100029 |
| | | 电子信箱(E-mail): | jly@bjjl.cn |

第 1 页共 4 页
Page 1 of 4

北京市计量检测科学研究院校准证书
Calibration Certificate of BIM

证书编号:
Certificate No. DA21S-AE100043

第 4 页共 4 页
Page 4 of 4

校准结果
Results of Calibration

2.扫描箱示值:(输入标准应变值2000με)

| 序号 | 显示值(με) | | 序号 | 显示值(με) | |
|---|---|---|---|---|---|
| | + | - | | + | - |
| 7-10-3 | 2000.8 | 2000.8 | 7-10-4 | 2000.8 | 2000.8 |

3.稳定性:

零点漂移: 0.4 με/4h

示值稳定度: 0.01 %/4h

测量结果不确定度: $U$=(0.05%×显示值+3)με ($k$=2)

注:本次测量设备示值符合JJG 623-2005电阻应变仪检定规程0.1级技术要求。

以下空白(End)

未经本院书面批准,不得部分复制本证书。
This certificate shall not be reproduced except in full, without written approval of BIM

**Fig. 5.** The Calibration-certificate of Beijing Institute of metrology

## 5.2 Field Test

In order to further verify the field application of the wireless strain testing system, the field testers took the instrument to Changqing oil field and Daqing oil field to test the stress and strain of oil Derrick. Different loads are applied to the Derrick, and the strain test results are shown in Table 4.

Table 4 shows the stress measurement values of channel 1 and channel 2 of No. 6 base station in the 19th set of wireless strain tester when different loads are applied to

**Table 4.** Strain test results of oil Derrick under different loads

| Strain (Mpa) | Load (KN) | | | |
|---|---|---|---|---|
| | 160 | 370 | 500 | 720 |
| STA-19-6-1 | 1.30 | −33.50 | −58.50 | −94.90 |
| STA-19-6-2 | 1 | −40.9 | −69.4 | −115.6 |

the Derrick. The corresponding data are imported into the Excel table to get the stress measurement value of the Derrick. The relationship between load and fitting curve, fitting curve and linear relationship are shown in Fig. 6. Through the analysis of the data in Table 1 and Fig. 6, it can be seen that the stress of the channel 1 test point of the No. 6 base station and the channel 2 test point of the No. 6 base station on the Derrick increases with the increase of load. Among them, the stress value of the channel 1 test point of the No. 6 base station has an ideal linear relationship with the applied load value (the fitting degree is 99.93%). The relationship between the stress value of STA-19-6-1 and the applied load value can be expressed as: y (stress) = − 0.1728 (load) + 29.215. The stress value of the 2-channel test point of the No. 6 base station has an ideal linear relationship with the applied load value (the degree of fitting is 99.97%). The relationship between the stress value of STA-19-6-2 and the applied load value can be expressed as: y (stress) = − 0.2088 (load) + 35.13.

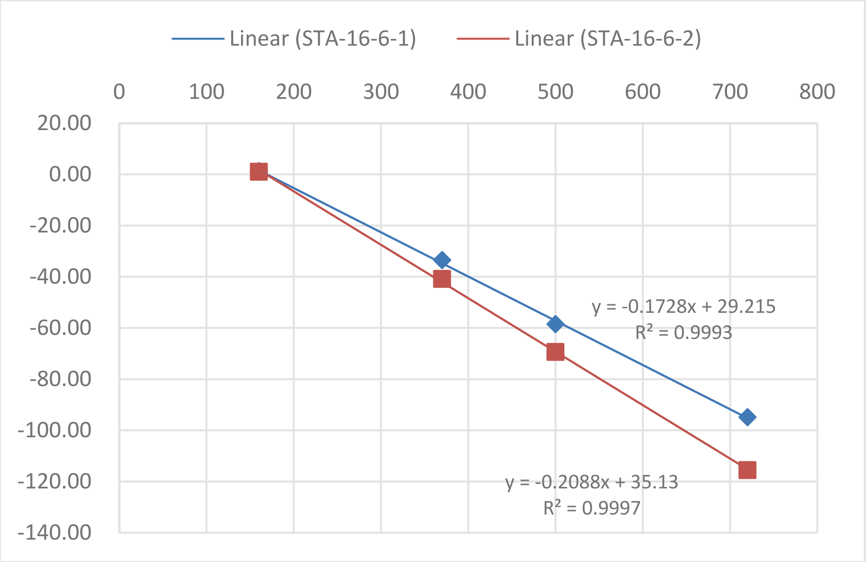
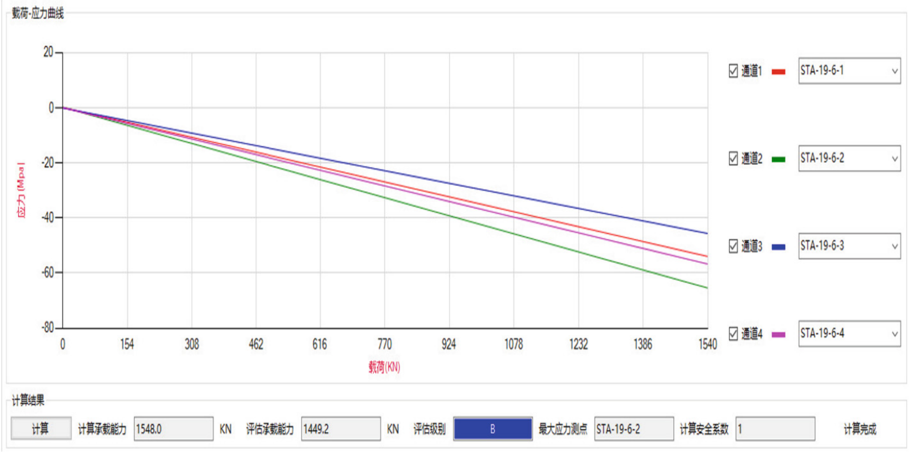


**Fig. 6.** Corresponding relation diagram of load and Derrick strain measurement

It can be seen from Fig. 6 that when different loads are applied to the derrick, the stress value measured by the wireless strain tester and the applied load value are in linear correspondence, which shows that the wireless strain tester and the wireless strain testing system perform strain state test when different loads are applied to the derrick, the system operates stably, the test data is accurate and reliable, and the degree of linear fitting is high.

According to the classification criterion of comprehensive evaluation method of Derrick comprehensive evaluation method in "SY6326–2019 oil drilling rig and workover rig Derrick base bearing capacity detection and classification standard", when the Derrick test hook load is greater than or equal to 95% of the designed maximum hook load, the Derrick evaluation grade is grade A; when the Derrick test hook load is less than 95% of the maximum hook load and greater than or equal to 85% of the designed maximum hook load, the Derrick evaluation grade is grade B. When it is less than 85% of the maximum hook load and greater than or equal to 70% of the designed maximum hook load, the Derrick evaluation level is level C. As can be seen from Fig. 7, according to the calculation of the load-bearing capacity of the Derrick in the wireless strain testing system, the calculated load-bearing capacity (ideal value) of the Derrick is 1548.0KN, the evaluated load-bearing capacity (actual value) is 1449.2KN, and the maximum capacity test point is STA-19-6-2 (that is, the maximum measured value of channel 2 of the No. 6 base station in the 19th set of wireless strain tester). After calculation, 1449.2/1548 = 93.62%, so the evaluation grade of derrick is B.



☑ 通道1 🟥 STA-19-6-1

☑ 通道2 🟩 STA-19-6-2

☑ 通道3 🟦 STA-19-6-3

☑ 通道4 🟪 STA-19-6-4

计算承载能力 1548.0 KN 评估承载能力 1449.2 KN 评估级别 B 最大应力测点 STA-19-6-2 计算安全系数 1 计算完成

**Fig. 7.** Evaluation grade calculation of bearing capacity of oil Derrick

The wireless strain test system can effectively reduce the labor intensity of testers because it saves the work of laying a large number of cables on the derrick; In addition, strain test and derrick bearing capacity calculation are integrated to improve the detection efficiency of oil derrick. Figure 8 and Fig. 9 are field inspection drawings. Compared with the traditional strain gauge, this instrument combined with wireless network transmission technology has the advantages of low power consumption, low cost, multiple nodes, high efficiency and so on. The test results and actual use show that the wireless strain test system has good engineering value.

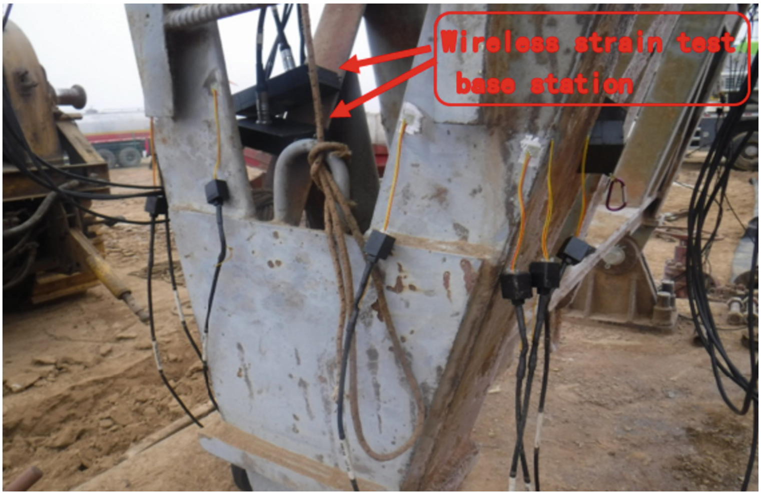


**Fig. 8.** Field test diagram 1 of wireless strain test system

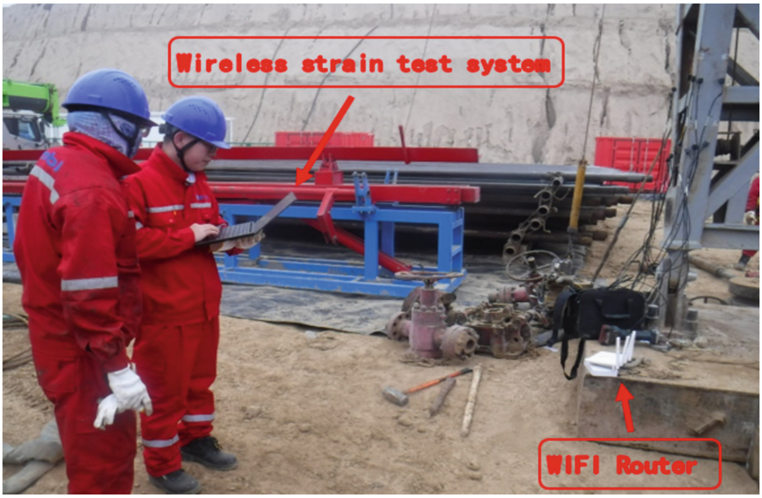


**Fig. 9.** Field test diagram 2 of wireless strain test system

## 6 Conclusion

The wireless strain test system is based on the strength theory of steel structure, and based on the precise strain measurement method and reliable wireless transmission technology. It integrates the strain test with the calculation of the derrick bearing capacity. The system solves the shortcomings of traditional instruments in field operation, such as poor reliability, heavy equipment, and high labor intensity, and has the ability to analyze the main bearing parts of drilling and workover equipment in real time. The system has high accuracy, reaching the highest accuracy level of the national resistance strain gauge metrological verification regulation; The system is targeted and can save a lot of working time. According to the statistics and analysis of the on-site testing work, the wireless strain testing system has the following advantages over the traditional stress and strain testing equipment:

1) Compared with wired instruments, the workload in the test phase is reduced by half
2) Compared with the general wireless test instrument, in the data acquisition and analysis stage, the influence of human factors on the data is eliminated, and the calculation results are more objective; Compared with the traditional instrument, it can export data to a fixed format, fill in test parameters, and calculate the bearing capacity manually, which improves the work efficiency by 2 times.
3) The field operation has good anti-interference and high reliability, and the measurement accuracy reaches the highest accuracy level of the national resistance strain gauge metrology verification regulation.

The wireless strain testing system has been widely used in the drilling field. It has detected and evaluated more than 6000 sets of drilling and workover rigs in major oilfields such as Daqing Oilfield, Changqing Oilfield, Xinjiang Oilfield, Dagang Oilfield, and has created direct economic benefits of tens of millions of yuan for the unit, and has played an important role in promoting scientific and technological progress and application of achievements in the industry.

**Acknowledgments.** The project is supported by Scientific Research and Technological Development Project of CNPC (Number 2021DJ4304).

## References

1. Cheng, L., Feng, G., Chuanxi, Z., et al.: Analysis of bearing capacity and residual life of offshore derrick with defects. Mach. Tool Hydraulics **50**(16), 124–129 (2022)
2. Deng, J., Xueli, L., Yi, Z., et al.: Monitoring system of derrick and substructure bearing capacity based on optical fiber sensing. Electron. Measure. Technol. **45**(10), 140–147 (2022)
3. Han, W.: Online detection of derrick capacity and hook load monitoring. China Pet. Mach. **48**(12), 23–26 (2020)
4. Dan, Z.: Research and Design of Safety Assessment and testing System for Derrick Capacity of Oil Drilling Rig, pp. 22–27. Lanzhou Jiaotong University, Lan Zhou (2019)
5. Miao, W.: Research on the Evaluation Method of the Derrick Bearing Capacity and Structure Strength of Workover Rig, pp. 44–48. Chong Qing Jiaotong University, Chong Qing (2020)
6. Han, W.: Research on Real-time Monitoring Technology of Derrick Bearing Capacity of Workover Rig, pp. 38–47. Southwest Petroleum University, Cheng Du (2017)
7. Yuhai, C.: Application rearch on evaluation method of defective workover rig derrick. Chem. Manage. **22**, 181–183 (2021)
8. Wang, L., Shan, P., Zhang, Y.: Design of High-precision multi-channel temperature and strain measurement system. Instrument Tech. Sens. **463**(08), 50–54 2021
9. Tao, L.: Research on Key Technology of Real—time Detection and Hooking Monitoring of Derrick Load Capacity, pp. 23–30. Yangtze University, Jing zhou (2018)
10. Nengyuan, T.: Design and Development of oil Derrick Structure Health Monitoring System, pp. 8–12. Yanshan University, Qin Huangdao (2019)

11. Zhang, S., Zhang, C., Wu, Q., et al. Research on the evaluation and classification of the bearing capacity of offshore workover rig derrick. China Pet. Mach. 34(7),47–51 (2018)
12. Xu Qiang, S., JianQing, L., et al.: Design of multi-channel multi-point wireless bridge strain monitoring system based on ZigBee. Transducer Microsyst. Technol. **39**(6), 79–82 (2020)

# A Method for Prediction of In-situ Stress Based on Empirical Formula and BP Neural Network

Chuan-gang Xiang[1](✉), Bo Chi[1], and Shu-yan Sun[2]

[1] Exploration and Development Research Institute of Da Qing Oilfield Co. LTD, Daqing, China
{xiangchuangang,chibo}@petrochina.com.cn
[2] Daqing Oilfield Drilling Engineering Company, Daqing, China
sunshuyan@petrochina.com.cn

**Abstract.** To solve the problems of complex in-situ stress of tight sandstone reservoir, few sample points of experimental data, difficulty in in-situ stress prediction, etc., a method for one-dimensional, two-dimensional and three-dimensional in-situ stress prediction based on geomechanics and BP neural network was innovatively proposed by comprehensively using various data such as core data, mechanical experimental data, logging data, etc. In this method, the rock mechanics parameters of single well in the study area were predicted by neural network method using the logging data as the learning sample and measured rock physical parameters as the monitoring data first; then the in-situ stress of single well was accordingly calculated by empirical formula, and predicted and analyzed by neural network algorithm using the calculated in-situ stress of single well selected by error analysis and the indoor measured in-situ stress as the monitoring data and the conventional logging data as the learning samples. The application in the actual areas shows that the predicted results of in-situ stress not only conform to the measured data, but also follow the logging curves, and thus provide an important basis for the design of integrated geological engineering scheme.

**Keywords:** BP neural network · error analysis · sample expansion · geostress prediction · Integration of geology and engineering

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20–22 September 2023.

This paper was selected for presentation by the IFEDC Committee following review of information contained in an abstract submitted by the author(s). Contents of the paper, as presented, have not been reviewed by the IFEDC Technical Team and are subject to correction by the author(s). The material does not necessarily reflect any position of the IFEDC Technical Committee its members. Papers presented at the Conference are subject to publication review by Professional Team of IFEDC Technical Committee. Electronic reproduction, distribution, or storage of any part of this paper for commercial purposes without the written consent of IFEDC Organizing Committee is prohibited. Permission to reproduce in print is restricted to an abstract of not more than 300 words; illustrations may not be copied. The abstract must contain conspicuous acknowledgment of IFEDC. Contact email: paper@ifedc.org.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 485–497, 2024.
https://doi.org/10.1007/978-981-97-0272-5_41

## 1 Introduction

With increasing difficulty in conventional energy exploitation, tight sandstone reservoirs have gradually become one of the important exploration and development targets [1–3]. Due to their low porosity and low permeability, they are mainly developed by horizontal drilling and staged fracturing, for which the current in-situ stress state is the main controlling factor [4–6]. In fracturing, the in-situ stress state controls the shape, height, width and direction of hydraulic fractures, and affects the fracturing stimulation effect. Also, in-situ stress is an important basis for well pattern deployment and adjustment and lateral segment direction selection. Therefore, the evaluation of the current in-situ stress is very important for the exploitation of tight sandstone reservoirs [7, 8].

Stress activity causes rock deformation or fracture and plays an important role in oil and gas exploration and development (Jenkins, 2017). The in-situ stress is generally composed of tectonic stress, gravity stress, thermal stress, pore pressure, etc., and its state is usually represented by three normal stresses, the maximum horizontal principal stress (σ H), vertical stress (σ v) and minimum horizontal principal stress (σ h) (Lin et al., 2006; Kuuskraa and Ammer. 2004; Matsuki and Takeuchi. 1993) [9, 10]. The magnitude and direction of in-situ stress in different strata in different regions in the crust change with space and time to form an in-situ stress field. Nowadays, there are many methods for evaluation of in-situ stress, such as hydraulic fracturing, acoustic emission and logging calculation. The methods are commonly used to determine the in-situ stress. Fracturing method refers to the calculation of the in-situ stress according to the relationship of fracturing pressure and time. The hydraulic fracturing method is reliable, direct and simple, but not all wells have small-scale fracturing test data in actual production. Acoustic emission method (Goodman 1963) refers to the determination of in-situ stress according to the characteristics of elastic waves in the form of which some energy is released from the rocks when the rocks are compressed by external forces [11]. By analyzing the acoustic characteristics of these elastic waves, the size of in-situ stress can be obtained. The acoustic emission method can measure the stress in the deep reservoirs, but has high costs of acoustic emission test, uncontinuity of results, and few sample points of indoor measured data and roughly estimates the in-situ stress state of the untested reservoirs. The logging curve-based formulas method overcomes the shortcomings of high test price and discontinuous measuring points. It refers to the method of indirect calculation of in-situ stress according to the formula after the logging data are used to calculate the rock mechanics parameters, such as formation Poisson's ratio, Young's modulus, shear modulus and bulk modulus, etc. However, the calculation process cannot be directly calibrated by the indoor experimentally measured in-situ stress results [12], and there is a large error between final calculation results and the measured data always [13, 14]. The neural network method better overcomes the disadvantage that the indoor experimentally measured in-situ stress results cannot be used for logging calibration, and seeks the relationship between the indoor experimentally measured in-situ stress results and the geophysical logging curve by machine learning. However, the relationship will be supported by a large number of the indoor experimentally measured in-situ stress results [15].

To solve the above problems, the data point continuity of logging curve-based formulas method and the advantages of data calibration of BP neural network method were combined to further improve the accuracy of in-situ stress prediction.

It includes the following steps: (1) The geophysical logging parameters of tight sandstone oil and gas wells, X-MAC logging parameters, indoor experimentally measured rock parameters and stress value parameters were acquired; (2) The standard layer statistical analysis method was used for environmental correction, and the histogram method was used to standardize the data of tight sandstone oil and gas wells; (3) The location correction method was used to correct the depth data used during the acquisition of the above data; (4) Linear regression method was used to predict the data of S-wave curve according to X-MAC logging parameters and processed geophysical logging parameters of the area; (5) According to the predicted S-wave slowness, geophysically logged P-wave slowness and density logging parameters, the dynamic Young's modulus and dynamic Poisson's ratio of rock were calculated jointly; according to the indoor experimental data, the rock mechanics parameters calculated according to the logging data were dynamically and statically corrected to obtain static Young's modulus, static Poisson's ratio, etc.; (6) according to the indoor experimentally measured data, the biot coefficient and pore pressure were obtained by fitting, and combined with the static rock mechanics parameters to calculate the vertical stress, maximum horizontal stress and minimum horizontal stress; and the error of the indoor experimentally measured stress data was analyzed; (7) the calculated in-situ stress of the wells with small errors was used to extend the indoor experimentally measured in-situ stress data of small samples to form large sample learning data; and the geophysical logging parameters of single wells were used as training samples to calculate the vertical in-situ stress, maximum horizontal stress and minimum horizontal stress of tight sandstone oil and gas wells by neural network method; (8) a geological mesh model was created to spatially interpolate geophysical logging parameters, and a three-dimensional in-situ stress field model was obtained based on neural network method to predict, analyze and evaluate the spatial in-situ stress distribution of tight sandstone reservoirs.

## 2 Sample and Experiment

A tight sandstone reservoir in Sanzhao Sag in the north of Songliao Basin was taken as an example. It has large depth, strong diagenesis, relatively tight lithology, active porosity of 13.5%, and air permeability of 0.9 × 10−3 $\mu m^2$. The natural gamma-ray logging curve of the whole well section shows "sandstone in mudstone", and the natural gamma-ray logging curve shows typical "sandstone-mudstone formation". The channel fill deposit and delta front deposit, especially anastomosing river flooding basin, mainly develop. The natural gamma-ray logging curves mainly show a dentate clock with medium/high amplitude, box, and small sawtooth with low amplitude. The data used were the logging data of 5 wells (G1, G2, G3, G4, and G5) in the area. There were 8 logging curves in total, including natural gamma-ray logging curve, natural potential curve, deep lateral resistivity curve, deep lateral resistivity curve, apparent resistivity curve, borehole diameter, density curve, and P-wave slowness curve. The reservoir more than 3550 m deep from the surface to underground was sampled at the interval of 0.125 m. In addition, G2 well has X-MAC logging data, including S-wave slowness, P-wave slowness, etc.

The standard layer statistical analysis method was used for environmental correction. The relationship between the target curve and the reference curve of the standard layer was statistically analyzed by using one or more logging curves less affected by the borehole as the generating curves (reference curves) and the curve to be corrected as the target curve. The correlation between the generating curve and the target curve was established for the non diameter-expanded section to predict the target curve of the trans-well section to be corrected. The multiple regression analysis method was used to establish the functional relationship. The coefficients of each reference curve were obtained by solving the equation, and then the coefficients and reference curves were used to create a mathematical model. This method, which is to apply the mathematical model to obtain new curves in the borehole diameter-expanded section, comprises the following steps: (1) the data of the depth section beyond the target layer (30 m away from top or bottom) were removed, and only the data of the target layer of tight sandstone reservoir were remained for later data processing; (2) the lateral profiles of geophysical logging curves, such as diameter curve, gamma ray curve, P-wave slowness curve and density curve of 5 wells, were plotted; (3) the wells or depth sections with abnormal borehole diameters were found out through horizontal comparison, and the influences of borehole diameters on different curves were analyzed; and (4) and other curves of the depth sections with large error at the expanded diameter were reconstructed by empirical formula or linear relationship using one or more logging curves less affected by the borehole as the generating curve.

The linear correlation analysis based on the data of other wells except G5 shows that the relationship between the logarithm of P-wave slowness and the logarithm of density is not very linear, while the relationship between the logarithm of P-wave slowness and the logarithm of natural gamma-ray logging data is very linear. The P-wave slowness curve was corrected for the diameter-expanded section of G5 well by the established linear relationship.

Histogram method was used for standardization. The frequency histogram of each logging curve in the standard layer of each well in the study area was plotted first, and then the histogram of other wells was compared with the histogram of a coring well or specific well as the standard histogram. If both of them have the same value and similar shape, their scales are accurate. If they are very different and their standard layers have consistent lithology, their scales are inaccurate, and the difference in response value between their standard layer and the standard layer of the coring well is the correction value.

In the area, 5 wells were tested by uniaxial compression and triaxial rock mechanics experiment. The depth at which the indoor experimentally measured data were acquired was corrected by the location correction method. In this method, a bar chart was plotted based on the indoor experimentally measured data and compared with the trend of the logging curve at the same depth. In case of the best agreement, the upward and downward movement distance of the core was the corrected core location value; a bar chart was plotted based on the measured core density and compared with the logging curve at the same depth. In case of the best agreement, the core movement was taken as the corrected location value. The location correction method can show the location depth difference intuitively. According to the above method, the indoor experimentally measured depth of the cores from 5 coring wells in the area was corrected for depth location, with the adjustment rate of 2.33%. The depth was not greatly corrected, with the average of 0.5 m and maximum of 1.2 m, within the allowable error range.

As shown in Fig. 2, the S-wave data calculation formula of tight sandstone in this area was established by linear regression between the S-wave slowness curve and the P-wave slowness curve of G2 well undergoing the X-MAC logging of Q1–Q7 layers. The regression equation has the calculation error of R2 = 0.897, indicating high correlation, so it can be used as a reference for calculation of the S-wave slowness curve data of other wells without X-MAC logging data. The S-wave slowness is calculated according to the following formula:

$$\Delta \mathrm{t_s} = 51.142 \times \mathrm{e}^{0.0121 \times \Delta \mathrm{t_p}} \tag{1}$$

where △ts and △tp are the S-wave slowness and P-wave slowness, respectively.

The data of the S-wave slowness curve of Q8–Q9 layers of G2 well were used as posterior data for error analysis. The results show that the S-wave predicted by linear regression is highly consistent with the measured S-wave, with the relative error of predicted S-wave slowness within 2%, indicating that the data of the S-wave slowness curve are so reliable as to provide a basis for calculation of tight sandstone reservoir rock parameters in the next step.

According to the different methods for their acquisition, the rock mechanics parameters are divided into two types, static parameters, which are obtained from uniaxial or triaxial loading tests of rock samples in a laboratory, and dynamic parameters, which are the mechanical parameters under various dynamic loads or periodically varying loads (such as acoustic, impact, vibration, etc.) calculated according to the data of the logging curve.

According to P-wave and S-wave propagation equations, the theoretical relationship between P-wave and S-wave velocity and dynamic rock parameters was given; the P-wave slowness Δtp and S-wave slowness Δts was obtained from the logging data; and the bulk density ρ was obtained from the density logging data in order to calculate various rock mechanics parameters, including dynamic Young's modulus E, bulk modulus K, shear modulus G, and dynamic Poisson's ratio μ.

The e dynamic Young's modulus E is calculated according to the following formula:

$$E = \frac{\rho}{\Delta t_s^2} \frac{3\Delta t_s^2 - 4\Delta t_p^2}{\Delta t_s^2 - \Delta t_p^2} \tag{2}$$

where ρ represents the bulk density.

The dynamic Poisson's ratio, μ, is calculated according to the following formula:

$$\mu = \frac{1}{2} \frac{\Delta t_s^2 - 2\Delta t_p^2}{\Delta t_s^2 - \Delta t_p^2} \tag{3}$$

The shear modulus, G, is calculated according to the following formula:

$$G = \frac{\rho}{\Delta t_s^2} = \frac{E}{2(1 + \mu)} \tag{4}$$

The bulk modulus, K, is calculated according to the following formula:

$$K = \rho \frac{3\Delta t_s^2 - 4\Delta t_p^2}{3\Delta t_s^2 \Delta t_p^2} = \frac{E}{3(1 - 2\mu)} \tag{5}$$

The deformation and rupture process of rock is slow and static. Before analysis of all rock mechanics, it is necessary to determine static parameters in order to predict the deformation and rupture of rock.

The elastic modulus directly calculated according to acoustic wave and density data is dynamic and thus impossible to predict the static mechanical properties of rocks. It is necessary to use a conversion formula obtained by laboratory data analysis to convert (generally linearly) the dynamic elastic modulus into static modulus, as shown in Fig. 3, so as to dynamically and statically correct the dynamic rock mechanics parameters calculated according to the logging data. According to the corrected static rock parameter results, the rock parameter profile of single well was plotted. It can be seen from Fig. 4 that the calculated static Young's modulus of G1 well is similar to the acoustic logging curve and density logging curve, and the static Poisson's ratio curve is negatively correlated to the acoustic logging curve and density logging curve, consistent with the trend of indoor experimentally measured data, indicating that the calculation result reflects the rock mechanical properties of the reservoir very well. According to the calculation results, the rock parameters of each single well in the study area can be evaluated.

The vertical stress, maximum horizontal stress and minimum horizontal stress of 5 wells in the study area were calculated according to the data of the density logging curve and static rock parameters. The number of experimentally measured in-situ stress values is too limited to obtain a continuous in-situ stress profile. However, the logging data have good continuity and high resolution, so it is very easy to popularize the method for prediction by empirical formula and BP neural network.

First of all, the data of the density logging curve were used to calculate the integral of formation thickness and density and further obtain the vertical in-situ stress. The calculation formula is as follows:

$$\sigma_{\mathrm{v}} = \int_0^{\mathrm{h}} \rho(h) g \cdot dh \tag{6}$$

where σv is vertical in-situ stress, MPa; h is the depth, m; ρ(h) is the density logging value, g/cm$^3$; g is the acceleration of gravity, m/s$^2$.

Then the formation pore pressure was calculated. In this embodiment, it was obtained by fitting the relationship between pressure and depth according to the indoor experimentally measured data.

Next, the biot coefficient was calculated. In this embodiment, it was obtained by fitting the indoor experimentally measured data. The indoor experimentally measured mineral volume content and porosity were taken as the input conditions to solve the biot coefficient (α), which is the ratio of formation pore space deformation to total volume change. The biot coefficient (α) is calculated according to the following formula:

$$\alpha = 1 - K_{\mathrm{dry}}/K_m \tag{7}$$

where $K_{dry}$ is the bulk modulus of dry rock, GPa; $K_m$ is the bulk modulus of mineral, GPa.

Then the biot coefficient was linearly fitted with the active porosity to obtain the expression of the biot coefficient of tight sandstone reservoir:

$$\alpha = 0.386 \cdot \ln(\varphi_{\text{eff}}) + 1.743 \tag{8}$$

where α is biot coefficient, dimensionless; Φeff is the active porosity of tight sandstone, %. The active porosity was calculated according to the logging curve in order to calculate the biot coefficients of single wells in the whole area and modeled work area.

Finally, the maximum horizontal in-situ stress and minimum horizontal in-situ stress were calculated according to the following formula based on the rock parameters, biot coefficient and pore pressure:

$$\begin{aligned} \sigma_{\text{h}} &= \frac{\nu}{1-\nu}(\sigma_{\text{v}} - \alpha P_{\text{p}}) + \frac{E\xi_{\text{h}}}{1-\nu^2} + \frac{\nu E\xi_{\text{H}}}{1-\nu^2} + \alpha P_{\text{P}} \\ \sigma_{\text{H}} &= \frac{\nu}{1-\nu}(\sigma_{\text{v}} - \alpha P_{\text{p}}) + \frac{E\xi_{\text{H}}}{1-\nu^2} + \frac{\nu E\xi_{\text{h}}}{1-\nu^2} + \alpha P_{\text{P}} \end{aligned} \tag{9}$$

where σH is the maximum horizontal principal stress, MPa; σh is the minimum horizontal principal stress, MPa; υ is Poisson's ratio, dimensionless; σv is the overburden pressure, MPa; α is biot coefficient, dimensionless; Pp is pore pressure, MPa; E is Young's modulus, GPa; εh and εH are the strain in the direction of minimum and maximum horizontal principal stress obtained by fitting the experimental data, respectively, dimensionless.

According to the measure in-situ stress of the tight sandstone reservoir in the north of Songliao Basin, the error of the calculated in-situ stress was analyzed. The error is the ratio of the difference between the calculated stress and the indoor experimentally measured stress to the measured stress. The error analysis shows that the overall error of vertical stress is within 5%, meeting the error requirement. However, the calculated minimum horizontal stress of G2, G4 and G5 wells and the calculated maximum horizontal stress of G5 well have large errors (more than 5%). Thus, the calculation of in-situ stress is a complex technical task. At present, there is no calculation formula suitable for all wells in one area. These calculation formulas are always called in-situ stress calculation models. The results calculated according to the formulas deviate from the actual in-situ stress value and thus will be corrected by other methods (Table 1).

In order to avoid the deviation between the calculated maximum and minimum horizontal stress, the implicit relationship between the in-situ stress of the reservoir and the conventional logging data was established by using the calibration and prediction advantages of BP neural network in order to form a new in-situ stress prediction model in the area. The correlation analysis between the measured stress data points of the core in the area and the conventional logging curves (P-wave slowness, neutron logging curve, density logging curve, gamma-ray logging curve) and depth attributes show that the total correlation coefficient between the sample data and the input layer is more than 91%. Thus, these logging parameters can be used as learning sample data for machine learning.

The area has some problems, such as only 48 in-situ stress measurement data points, few learning samples, discontinuous data, multiple solutions, and random data between

**Table 1.** Error of Calculated In-situ Stress

| Well No. | Error of calculated in-situ stress | | | |
|---|---|---|---|---|
| | Depth range (m) | Relative error of vertical stress (%) | Relative error of maximum horizontal stress (%) | Relative error of minimum horizontal stress (%) |
| G1 | 1773–1875 | 0.63 | 2.82 | 2.87 |
| G2 | 2015–2100 | 0.30 | 3.29 | 5.24 |
| G3 | 2015–2110 | 2.24 | 2.04 | 3.27 |
| G4 | 1793–1886 | 0.83 | 3.75 | 5.02 |
| G5 | 1996–2141 | 1.13 | 8.05 | 7.63 |
| Average | - | 0.90 | 3.99 | 4.86 |

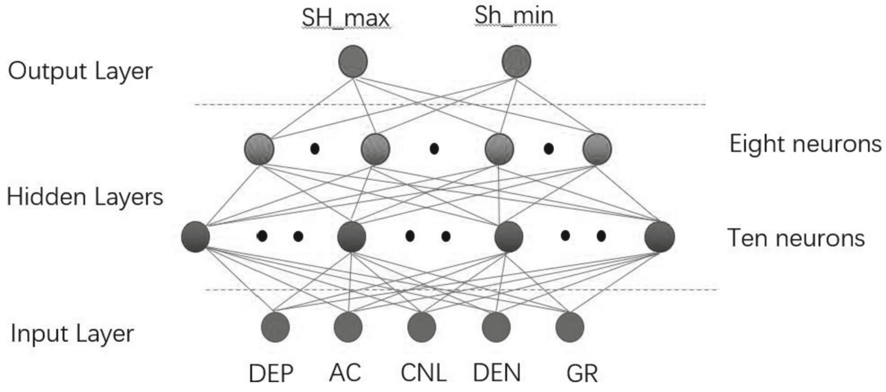


**Fig. 1.** Prediction of single well in-situ stress by BP neural network

measurement points. In order to solve the above problems, the neural network method calculation formula method were organically combined together, i.e. the continuous data of the wells with small errors in the calculated in-situ stress results were used as the sample data of machine learning, to extend the sample data from small core (discontinuous) samples to large logging (continuous) samples, further reduce the possibility of multiple prediction results and thus keep a good correlation between the prediction result and logging curve. Specifically, the continuous vertical stress, calculated maximum and minimum horizontal stress of G1 well and G3 well at different depths with the error of calculated stress value within 3% were selected to extend the data of 48 measured in-situ stress parameters (small samples) and thus form large sample (learning sample) data; the vertical in-situ stress, maximum horizontal stress and minimum horizontal stress of tight sandstone oil and gas wells were calculated by BP neural network model using the geophysical logging (acoustic logging, neutron logging, density logging, gamma-ray logging) curve of single well and depth value as training samples. After determination of the above training input parameters of machine learning, the BP neural network model (as shown in Fig. 7) consisting of one input layer, two hidden layers and one output layer built by matlab was used to select Levenberg-Marquardt backpropagation for training and predict the maximum and minimum horizontal principal stress of G2, G4 and G5 wells. Tables 5–6 are obtained through error statistics. The error analysis shows that the maximum horizontal principal stress and the minimum horizontal principal stress predicted by BP neural network based on the indoor experimentally measured data have the average error of 4.73% and 10.28%, respectively. Compared with the error of the calculated in-situ stress, the error doesn't reduce. The maximum and minimum horizontal principal stresses predicted by BP neural network based on the extended large sample have the average error of 1.54% and 1.45%, respectively. The prediction by BP neural network model shows that the extended large sample has much smaller error than the small sample. Therefore, the calculated in-situ stresses of the wells with small error were used to extend the learning sample of the indoor experimentally measured data and thus change the small sample into a large sample for prediction by BP neural network. Thus, the prediction result is more accurate. This method can be used to predict and comprehensively evaluate the in-situ stress of other wells in this area.

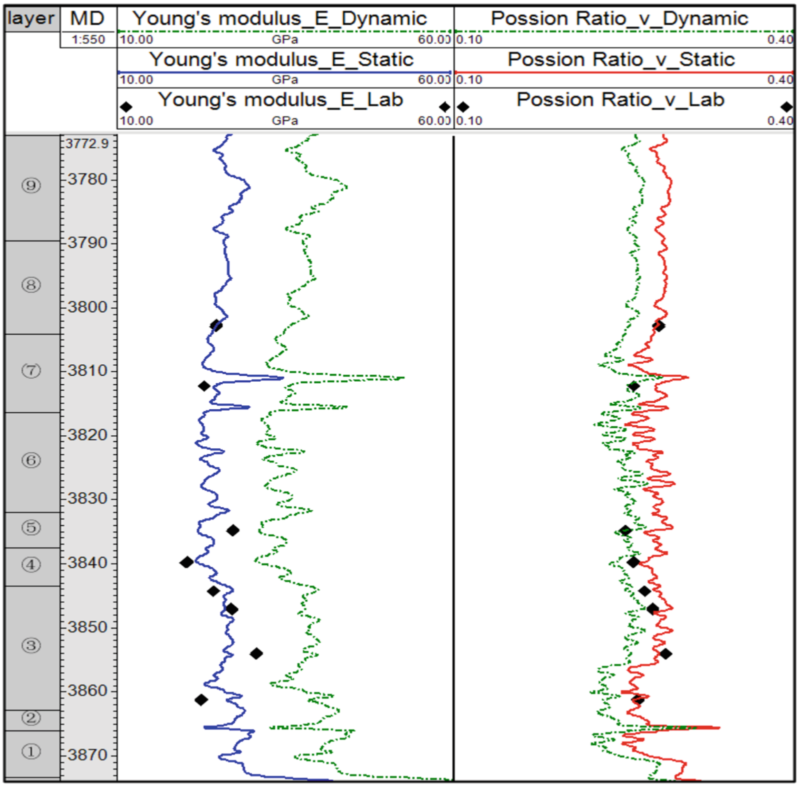


**Fig. 2.** Indoor measured and Calculation of dynamic-static Poisson's ratio or Young's modulus based on logging curve

**Table 2.** Comparison of errors of the in-situ stress predicted by neural network models based on different learning samples

| Well No. | Error of the in-situ stress predicted by formula method (%) | | Error of the in-situ stress predicted by BP neural network only based on indoor experimentally measured data (%) | | Error of the in-situ stress predicted by BP neural network based on extended large samples (%) | |
|---|---|---|---|---|---|---|
| | Maximum horizontal stress | Minimum horizontal stress | Maximum horizontal stress | Minimum horizontal stress | Maximum horizontal stress | Minimum horizontal stress |
| G2 | 3.29 | 5.24 | 5.23 | 12.07 | 0.16 | 0.21 |
| G4 | 3.75 | 5.02 | 4.54 | 10.86 | 3.61 | 3.34 |
| G5 | 8.05 | 7.63 | 4.42 | 7.93 | 0.85 | 0.80 |
| Average error | 5.03 | 5.96 | 4.73 | 10.28 | 1.54 | 1.45 |

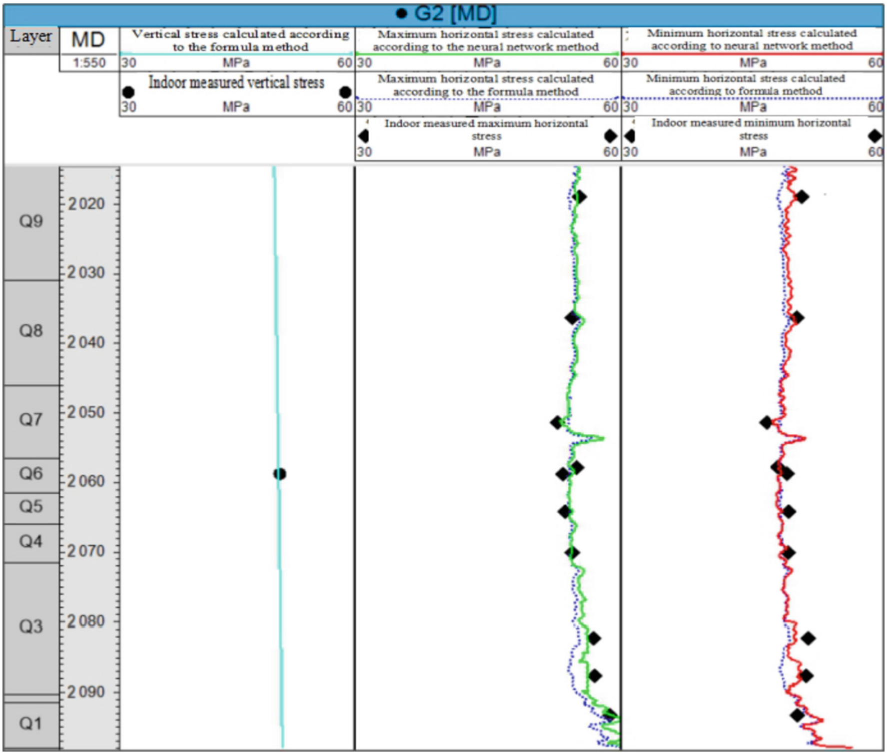
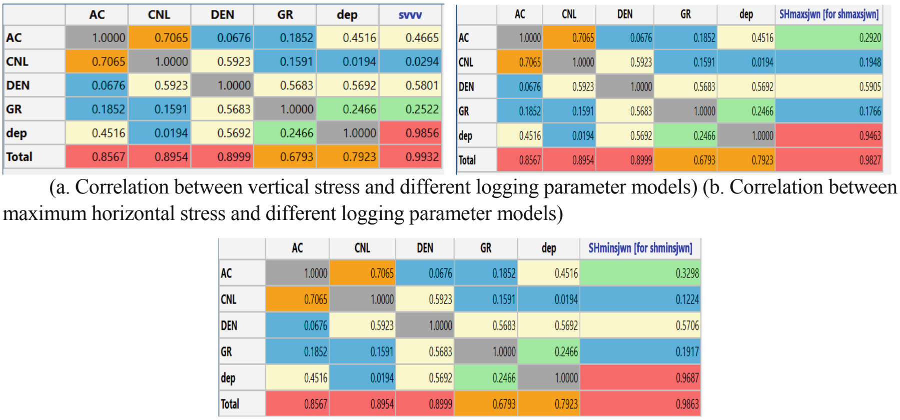


**Fig. 3.** Comparison between the calculated results of the three-dimensional stress neural network and the measured values

## 3 Results and Discussions

The above method can further extend the establishment of in-situ stress geological model, i.e. the fine geological model was created by using log and seismic data first and then the acoustic wave slowness curves, neutron logging curves, density logging curves, and gamma-ray logging curves of 5 wells in the area were discretized to obtain the averages of each layer in each well. The area has large area and only 5 wells, so the co-kriging interplotation algorithm was used for spatial interpolation to obtain the spatial distribution characteristics of different parameters.

The interpretation results of the above-mentioned in-situ stress were discretized to obtain the correlation between statistical data points and conventional logging curves (acoustic wave slowness logging curve, neutron logging curve, density logging curve, gamma-ray logging curve) and depth attribute model. The neural network algorithm shows that the total correlation coefficient exceeds 0.98. Therefore, it is feasible to use these logging parameter models as learning data and the measured in-situ stress as sample data for machine learning.



| | AC | CNL | DEN | GR | dep | svvv |
|---|---|---|---|---|---|---|
| AC | 1.0000 | 0.7065 | 0.0676 | 0.1852 | 0.4516 | 0.4665 |
| CNL | 0.7065 | 1.0000 | 0.5923 | 0.1591 | 0.0194 | 0.0294 |
| DEN | 0.0676 | 0.5923 | 1.0000 | 0.5683 | 0.5692 | 0.5801 |
| GR | 0.1852 | 0.1591 | 0.5683 | 1.0000 | 0.2466 | 0.2522 |
| dep | 0.4516 | 0.0194 | 0.5692 | 0.2466 | 1.0000 | 0.9856 |
| Total | 0.8567 | 0.8954 | 0.8999 | 0.6793 | 0.7923 | 0.9932 |

| | AC | CNL | DEN | GR | dep | SHmaxsjwn [for shmaxsjwn] |
|---|---|---|---|---|---|---|
| AC | 1.0000 | 0.7065 | 0.0676 | 0.1852 | 0.4516 | 0.2920 |
| CNL | 0.7065 | 1.0000 | 0.5923 | 0.1591 | 0.0194 | 0.1948 |
| DEN | 0.0676 | 0.5923 | 1.0000 | 0.5683 | 0.5692 | 0.5905 |
| GR | 0.1852 | 0.1591 | 0.5683 | 1.0000 | 0.2466 | 0.1766 |
| dep | 0.4516 | 0.0194 | 0.5692 | 0.2466 | 1.0000 | 0.9463 |
| Total | 0.8567 | 0.8954 | 0.8999 | 0.6793 | 0.7923 | 0.9827 |

(a. Correlation between vertical stress and different logging parameter models) (b. Correlation between maximum horizontal stress and different logging parameter models)

| | AC | CNL | DEN | GR | dep | SHminsjwn [for shminsjwn] |
|---|---|---|---|---|---|---|
| AC | 1.0000 | 0.7065 | 0.0676 | 0.1852 | 0.4516 | 0.3298 |
| CNL | 0.7065 | 1.0000 | 0.5923 | 0.1591 | 0.0194 | 0.1224 |
| DEN | 0.0676 | 0.5923 | 1.0000 | 0.5683 | 0.5692 | 0.5706 |
| GR | 0.1852 | 0.1591 | 0.5683 | 1.0000 | 0.2466 | 0.1917 |
| dep | 0.4516 | 0.0194 | 0.5692 | 0.2466 | 1.0000 | 0.9687 |
| Total | 0.8567 | 0.8954 | 0.8999 | 0.6793 | 0.7923 | 0.9863 |

(c. Correlation between minimum horizontal stress and different logging parameter models)

**Fig. 4.** Spatial Distribution Model of Logging Parameters

The models of vertical stress, maximum horizontal stress and minimum horizontal stress were created by neural network algorithm. The in-situ stress prediction shows that the study area has the characteristics of maximum horizontal principal stress > vertical principal stress > minimum horizontal principal stress, i.e. Class III in-situ stress in the state of sliding stress, and always form vertical fractures after fracturing. Therefore, the area can be developed by horizontal well + fracturing; the horizontal well trend is perpendicular to the direction of the maximum horizontal stress, so the north-north-east well spacing is suggested. The study on the stress difference distribution shows that due to relatively large stress difference in the axial region and anticline, intensive cut fracturing of horizontal wells is suitable for the large stress blocks to improve the producing degree; the basal leaf cross has smaller stress difference than the anticline, so large cluster distance fracturing is suitable for small stress blocks to form complex artificial fractures and expand the swept volume.

## 4 Conclusions

Compared with the existing technology, this method has the following beneficial effects: (1) By comprehensively using core analysis data, indoor measured mechanical data, conventional logging data, XMAC logging data, etc., the method for prediction of one-dimension, two-dimension and three-dimension in-situ stress prediction integrating artificial intelligence and traditional geomechanics was innovatively put forward based on research on structure, sedimentary facies and reservoir properties in order to extend small core samples to large logging samples; (2) by creating a neural network-based three-dimensional in-situ stress prediction model, the spatial in-situ stress distribution of tight sandstone reservoirs in the test area was predicted and analyzed. The relative error between the predicted in-situ stress and the measured result is within 3%, indicating that the method improves the prediction accuracy of in-situ stress in the study area. Thus, the research results provide an important basis for the design of integrated geological engineering scheme. In conclusion, this method is very worthy of application in stress prediction technology for tight sandstone reservoirs due to its advantages, including simple logic, accuracy and reliability.

## References

1. Boswell, L.F., Chen, Z.: A general failure criterion for plain concrete. Int. J. Solids Struct. **23**(5), 621–630 (1987)
2. Cai, W., Zhu, H., Liang, W., Zhang, L., Wu, W.: A new version of the generalized Zhang–Zhu strength criterion and a discussion on its smoothness and convexity. Rock Mech. Rock Eng. 1–17 (2021)
3. Colmenares, L.B., Zoback, M.D.: A statistical evaluation of intact rock failure criteria constrained by polyaxial test data for five different rocks. Int. J. Rock Mech. Min. Sci. **39**(6), 695–729 (2002)
4. Jiang, J., Pietruszczak, S.: Convexity of yield loci for pressure sensitive materials. Comput. Geotech. **5**(1), 51–63 (1988)
5. Kim, M.K., Lade, P.V.: Modelling rock strength in three dimensions. Int. J. Rock Mech. Min. Sci. Geomech. Abstracts **21**(1), 21–33. Pergamon (1984)
6. Lee, Y.K., Pietruszczak, S., Choi, B.H.: Failure criteria for rocks based on smooth approximations to Mohr-Coulomb and Hoek-Brown failure functions. Int. J. Rock Mech. Min. Sci. **56**, 146–160 (2012)
7. Li, C., Li, C., Zhao, R., Zhou, L.: A strength criterion for rocks. Mech. Mater. **154**(3), 1–9 (2021)
8. Lade, P.V., Duncan, J.M.: Elastoplastic stress-strain theory for cohesionless soil. J. Geotech. Eng. Div. **101**(10), 1037–1053 (1975)
9. Pan, X.D., Hudson, J.A.: A simplified three-dimensional Hoek-Brown yield criterion. In: ISRM International Symposium. OnePetro (1988)
10. Priest, S.D.: Determination of shear strength and three-dimensional yield strength for the Hoek-Brown criterion. Rock Mech. Rock Eng. **38**(4), 299–327 (2005)

11. Yu, M., Zan, Y., Xu, S.: Rock Strength Theory and Its Application. Science Press (2017). (in Chinese)
12. Zan, Y., Yu, M., Wang, S.: Nonlinear unified strength criterion of rock. Chin. J. Rock Mech. Eng. **21**(10), 1435–1441 (2002). (in Chinese)
13. Zan, Y., Yu, M., Zhao, J.: Nonlinear unified strength theory of rock under high stress state. Chin. J. Rock Mech. Eng. **23**(13), 2143–2148 (2004). (in Chinese)
14. Zan, Y., Yu, M.: Generalized Nonlinear Unified Strength Theory of Rock. J. Southwest Jiaotong Univ. **48**(4), 616–624 (2013). (in Chinese)
15. Zhang, Q., Zhu, H., Zhang, L.: Modification of a generalized three-dimensional Hoek-Brown strength criterion. Int. J. Rock Mech. Min. Sci. **59**, 80–96 (2013)

# Design and Implementation of A2 System Regional Center Data Synchronization Scheme

Yang Jiao[1,2](✉), Shan Xie[1,2], Hong-mei Deng[1,2], and Jian-hua Su[1,2]

[1] Research Institute of Exploration and Development, Changqing Oilfield Company, Xi'an, Shaanxi, China
jyang_cq@petrochina.com.cn

[2] National Engineering Laboratory for Exploration and Development of Low Permeability Oil and Gas Fields, Xi'an, Shaanxi, China

**Abstract.** Based on the production data management system of oil, gas and water wells (A2), Changqing Oilfield has developed comprehensive query, big data analysis and other deepening application functions, which play an important supporting role in oil and gas exploration deployment, production dynamic analysis and other work. At present, Changqing A2 Regional Center has many problems in daily production data synchronization, such as long synchronization time, low timeliness, weak verification logic, and data timeliness and accuracy are facing challenges. In view of the above difficulties, this paper designed a new data synchronization scheme and completed the feasibility test of key technical points by analyzing the publish and unlock process of A2 data and the current synchronization logic, and built a full process closed-loop synchronization system of "publish awareness, decentralized synchronization, unlock identification and local update". Through actual operation, the average data synchronization time was shortened from 83 min to 2 min, and the unlock recognition rate was improved to 100%, which realized the efficient application of A2 production data and boosted the high-quality secondary accelerated development of Changqing Oilfield. At the same time, the scheme has good guidance and reference significance for other oilfield companies that need to build A2 regional data center.

**Keywords:** A2 System Regional Center · Data synchronization · Unlock identification · Full process closed loop

## 1 Introduction

The production data management system for oil, gas and water wells (A2) is a professional information system built by China National Petroleum Corporation (CNPC), aiming at the organization and management of the core data of oil and gas field production and development. Changqing Oilfield has developed a series of comprehensive query, big data analysis and other in-depth application function modules based on A2, which play an important supporting role in oil and gas exploration deployment, development plan adjustment, production performance analysis and other work [1]. In 2018, the

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 498–504, 2024.
https://doi.org/10.1007/978-981-97-0272-5_42

A2 system of Changqing Oilfield was migrated from Xi'an to Beijing Changqing Data Center. After the migration, due to network bandwidth, link stability and other reasons, the A2 related oilfield self-built system cannot connect to the Beijing A2 database (Beijing database), resulting in the normal operation of each system. Therefore, Changqing Oilfield has redeployed a set of databases (Xi'an database) that are isomorphic to Beijing Database in Xi'an. At first, GoldenGate software was used for real-time data synchronization, but due to frequent failures, it brought huge workload of operation. Therefore, a programmed automatic data synchronization mode was adopted every night to support the data application of the oilfield self-built system. This synchronization mode does not judge A2 data modification, which makes it difficult to ensure data accuracy. In addition, the data published today can be used only the second day, which is difficult to meet the new needs of the secondary accelerated development of the oilfield. It is urgent to build a new data synchronization system [2].

## 2 Research on the Mechanism of A2 Data Synchronization and Unlocking

Currently, A2 data synchronization consists of three phases: company data publishing, basic entity synchronization and daily data synchronization. Company data publishing needs to go through three levels of auditing, including operation area, oil/gas production plant and oil field company. When the second level auditing is completed by the oil/gas production plant, A2 system will automatically summarize and check the data of every plants under the jurisdiction of Changqing Oilfield Company, and complete the first level auditing and the official publish of company data, usually before 15:00 a day. Basic entities, including well basic information, well bore information, well status history information, etc., start up at 20:00 a day and take an average of 3 min. Daily data synchronization involves common data such as production day data, mining machine data and test sample data of 36 plants. The process starts at 22:00 and takes an average of 83 min to complete deletion and increase of data in the last three days.

A2 data unlocking refers to the publish of company-level data after the completion of the publish, for the existing abnormal data, each plant can apply to unlock all the data this month from the date of the abnormal, the unlocked data can be modified, and need to publish and synchronize again. However, the current synchronization logic only complete deletion and increase of data in the last three days. It cannot accurately identify the unlocking plants and date range, and cannot process data for more than three days. It requires manual problem detection and manual day-by-day processing.

In summary, the existing A2 data synchronization mechanism has simple logic, less code, and artificial operation, which achieves the data "published today, applied the next day". However, with the increasing speed of scientific research and production, more and more oil/gas production plants put forward the urgent need for data "published today, applied today". It is urgent to build an agile synchronization system with high efficiency, accurate judgment and reliable data [3].

## 3 Data Synchronization Scheme Design

### 3.1 Overall Scheme Design

Based on the analysis of existing synchronization logic, the overall implementation scheme of A2 data synchronization designed in this paper includes basic entity synchronization, daily data synchronization and unlocked plant data resynchronization. The scheme can synchronize the data in time after the initial release of the data of each plant, and accurately identify and resynchronize the data unlock of the current month in the subsequent synchronization process [4, 5]. The scheme changes the data synchronization mode from synchronizing the data of 36 plants for 3 days at a time to real-time synchronizing the data published by each plant (See Fig. 1).

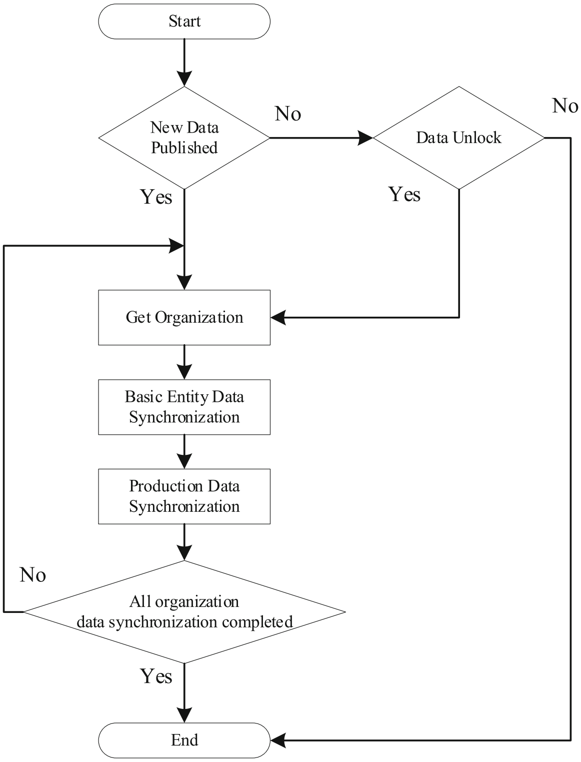

Start
New Data Published
No
Data Unlock
No
Yes
Yes
Get Organization
Basic Entity Data Synchronization
Production Data Synchronization
No
All organization data synchronization completed
Yes
End

**Fig. 1.** Overall scheme of data synchronization

### 3.2 Release Status Judgment of Oil/Gas Production Plant

The data audit at the oil/gas production plant level is a secondary audit. The audit shows that the production data to be synchronized on that day has been released at the plant level. Since the plant-level release status is a sufficient and unnecessary condition for each data table to complete the data entry on the current day, the data table containing the approval status field within the synchronization scope can be used as the starting data synchronization condition.

### 3.3 Unlock Status Judgment

The essence of unlocking is the re-publishing of data, which is completely opposite to the process of data publishing. If the changed data is in the operation area, it needs to be unlocked by the company, then by the plant level, and finally by the operation area, and the data will be reprocessed. For data unlocking over a period of time, it is necessary to unlock data day by day from the current data release date until the start date of the data change. Since data needs to be republished after unlocking, data unlocking can change the data release time. By comparing the data release time, it can be determined whether there is data unlocking or multi-day unlocking in the oil/gas production plant [6].

## 4 Construction of Whole-Process Synchronization System

### 4.1 Data Table Structure Analysis

In terms of daily data synchronization, the old delete all add all data synchronization method does not need to determine which company the data in the data table belongs to, but the new synchronization method needs to specify how the company relates to the data in different data tables, so that the data of a company in the table can be accurately and completely synchronized after the data of a company is published. Daily data synchronization involves more than 20 data tables, including well completion layer table, production layer status history table, daily data of injection well status, and daily data of production well status. The primary key is divided into two types: well ID and entity ID. Well ID is directly associated with the organization ID. Entity ID includes geological unit, combination unit and station library, and each type is associated with the organization ID. Through the above association relationship, the data to be synchronized in the data table can be accurately identified after the oil/gas production plant data is released [7].

### 4.2 Effective Field Analysis

A2 system is developed based on EPDM model, and there are a certain number of invalid fields in actual application. After testing, reducing the number of synchronized fields can improve the efficiency of data synchronization. Therefore, invalid fields need to be judged and filtered. First of all, the fields with the number of duplicate values less than 10 should be initially filtered, and then the total data volume of the above fields containing duplicate values, the data volume of non-null values, and the specific data of non-null values and other auxiliary information should be obtained. Finally, whether the field is invalid can be determined through manual judgment of the auxiliary information. For example, Data table X contains field Y, in which the number of non-repeating values is 1, the total data amount is 34452797, the total effective data amount is 128629, and the total effective field value is 0. It can be determined from the above information that this field is invalid.

### 4.3 Synchronous Log Construction

By comparing the data release time of Beijing Database with the data synchronization start time of Xi'an Database, the logical trend of the synchronization scheme can be obtained (see Fig. 2). Therefore, date and time are the key data that must be recorded. Before building the data synchronization logic, the data synchronization log table needs to be established first. The log table mainly includes six aspects: name of the oil/gas production plant, data release time, data synchronization start time, data synchronization end time, synchronization consumption time, and remarks. The data synchronization log table not only determines the direction of synchronization logic, but also records the synchronization and unlocking status of data in a timely manner to assist in the analysis of synchronization efficiency and troubleshooting of synchronization errors [8].

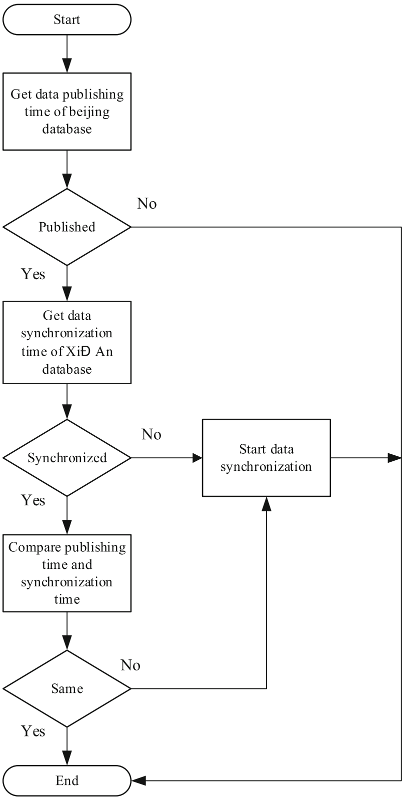

Start
Get data publishing time of beijing database
Published
No
Yes
Get data synchronization time of XiÐ An database
Synchronized
No
Start data synchronization
Yes
Compare publishing time and synchronization time
Same
No
Yes
End

**Fig. 2.** Data synchronization and unlocking judgment logic

### 4.4 Synchronization Logic Construction

The data synchronization logic of A2 regional center is mainly based on the data release time of Beijing database and the data synchronization time of Xi'an database. According to this scheme, data synchronization is started at 9:00 a.m. every day. First, the data synchronization oil/gas production plants are obtained and written to the data synchronization log table. Then, the oil/gas production plants in the data table are periodically traversed, and the data release and data synchronization time of each plants are queried circularly. According to the data synchronization and unlocking judgment logic in Fig. 2, the initial synchronization or unlocking resynchronization of data is started. After the synchronization of all oil/gas production plants are completed, continue to judge whether the data of the current day is unlocked until 11:00 p.m. At 10:45 p.m., a special judgment will be made. If the number of units completing data synchronization on the current day is less than the total number in the synchronization log data table, a full data synchronization of the current day will be started, and the operation will be recorded in the synchronization log data table. The system administrator will be notified the next day to check the data integrity. At 10:45 p.m., a special judgment will be made. If the number of oil/gas production plant completing data synchronization on the current day is less than the total number in the data synchronization log table, a full synchronization of the current day's data will be started, and the operation will be recorded in the data synchronization log table, and the system administrator will be notified to check the data integrity [9].

## 5 Application Effect

### 5.1 Improve the Data Unlocking Recognition Rate

Based on the statistical analysis of data synchronization logs, the average daily data synchronization time of each unit is about 120 s (see Fig. 3). Compared with the original data synchronization time of 83 min, the efficiency is increased by more than 20 times.

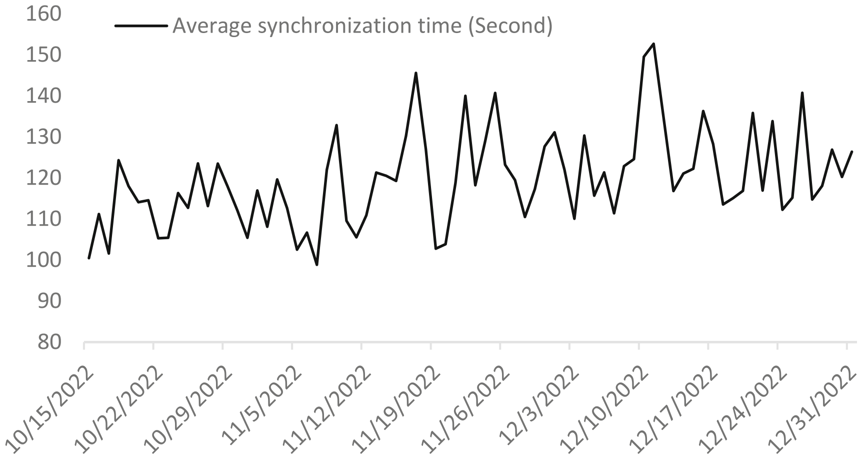


**Fig. 3.** Average synchronization time of new scheme

### 5.2 Improve the Accuracy of Data Unlocking

By comparing the data release time in reverse order between the Beijing database and the data synchronization log table, the accurate identification and resynchronization of cross-date data unlocking are realized, and the data accuracy and consistency are effectively improved.

## 6 Conclusion

The A2 system regional center data synchronization solution solves the problem that the oilfield self-built information system is difficult to directly apply A2 data. The core of the scheme is to identify the release status and unlock status of the data. At the same time, the efficiency of data synchronization has been improved again by simplifying the invalid fields, and finally the "published today, applied today" of A2 production data has been realized. It has good guidance and reference significance for other oilfield companies that need to build the regional center of A2 system, or the database cannot be directly connected to the application due to network, efficiency, synchronization software and other problems, the business scenario that needs to build a regional data center has good guidance and reference significance.

## References

1. Tao, F.: Establishment and promotion of an integrated application system for oil, gas, and water well production data. Commun. World **06**, 272–273 (2017)
2. Tian, B.: Research on the application of logical data synchronization technology based on oracle database in practice. Comput. Program. Skills Mainten. **15**, 31–32 (2016)
3. Wang, H., Mu, L., et al.: Optimized management and fast query of decentralized storage of oil and gas production dynamic big data. Pet. Explor. Dev. **46**(05), 959–965 (2019)
4. Kou, Y., Wang, X., et al.: Research and application of database synchronization technology. Comput. Knowl. Technol. **15**(15), 3–5+34 (2019)
5. Chen, J., Yan, J., et al:. An optimized database synchronization method and device. Guangxi Commun. Technol. (02), 1–4+9 (2017)
6. Wang, Z., Xiu, W.: Database synchronous programming solution. J. Liaoning Normal Univ. (Nat. Sci. Ed.) **19**(01), 32–33+58 (2017)
7. Cui, J.: Application of master data management system in enterprise informatization construction projects. Eng. Constr. Des. **21**, 201–203 (2021)
8. Wang, G., Wang, Y., et al.: Research on interface technology for synchronized reporting of A2 and development database. China Petrol. Chem. Stand. Qual. **34**(06), 165–166 (2014)
9. Xia, B.: Research and application of database synchronization based on oracle advanced replication function. Inf. Syst. Eng. **03**, 87 (2015)

# An Attention-Based Temporal and Spatial Convolution Recursive Neural Network for Surrogate Modeling of the Production Curve Prediction

Xu Chen, Kai Zhang(✉), Xiao-ya Wang, Jin-ding Zhang, and Li-ming Zhang

School of Petroleum Engineering, China University of Petroleum (East China), Qingdao, China
zhangkai@upc.edu.cn

**Abstract.** Reservoir numerical simulation is a time-consuming and expensive procedure, especially for production optimization and history matching, which requires multiple calls to the reservoir numerical simulator. The surrogate model based on deep learning can provide a proxy solution to the process of calculating the production curve by the numerical reservoir simulator with approximate accuracy and higher computational efficiency, while an attention mechanism can better capture the local time characteristics of production time series. The attention mechanism is introduced based on deep CNN and LSTM to build an attention-based temporal and spatial convolution recursive neural network for surrogate modeling which is capable of extracting spatial features of reservoir static parameters and handling temporal data, and establishing an image-sequence mapping relationship from reservoir static parameters to reservoir production curve which is used to predict the reservoir production curve. The constructed surrogate model can fast and accurately predict production curves and improves the computational efficiency of production optimization and history matching.

**Keywords:** Deep Learning · Surrogate Model · Production Curve Prediction · Attention Mechanism

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20–22 September 2023.
This paper was selected for presentation by the IFEDC Committee following review of information contained in an abstract submitted by the author(s). Contents of the paper, as presented, have not been reviewed by the IFEDC Technical Team and are subject to correction by the author(s). The material does not necessarily reflect any position of the IFEDC Technical Committee its members. Papers presented at the Conference are subject to publication review by Professional Team of IFEDC Technical Committee. Electronic reproduction, distribution, or storage of any part of this paper for commercial purposes without the written consent of IFEDC Organizing Committee is prohibited. Permission to reproduce in print is restricted to an abstract of not more than 300 words; illustrations may not be copied. The abstract must contain conspicuous acknowledgment of IFEDC. Contact email: paper@ifedc.org.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 505–515, 2024.
https://doi.org/10.1007/978-981-97-0272-5_43

## 1 Introduction

The intelligence of oilfields has become the mainstream trend of oilfield development at present, and the fine modeling of oilfields has become an essential guarantee for the real-time dynamic adjustment of production in intelligent oilfields [1]. The complex geological conditions and geological such as permeability and porosity make the numerical model with non-negligible uncertainties, resulting in a very challenging fine description of reservoir model properties [2, 3]. The automatic history matching method corrects the oil reservoir numerical simulation results to decrease uncertainty of oil reservoir parameters which aims to fit the oil reservoir historical observation results, so as to dynamically adjust the random geological model initialized according to the field measurement data of the oilfields to approach the real geological model which is a general fine-grained reservoir modeling method adopted in the field of petroleum engineering [4]. The automatic history matching process requires repeatedly drawing on the oil reservoir numerical simulator to characterize the reservoir fluid flow process and seepage characteristics and calculate the reservoir numerical model's oil and water production data. Because of the complexity of oilfield reservoir distribution and the complex nonlinear flow features of reservoir fluid flow, the oil reservoir numerical simulation is time-intensive and expensive resulting in the automatic history matching process becoming a typical high-dimensional, time-consuming, and expensive inverse problem-solving procedure [5]. Despite the existence of many ways and measures to improve computational efficiency, the technical difficulties associated with expensive and time-consuming calculations remain [6, 7].

Encouraged by the introduction of deep learning in petroleum engineering, the model, such as a deep neural network, can learn the hierarchical characteristics of the reservoir geological grid attributes similar to the picture pixel format, and establish the nonlinear regression relationship from image data to sequence data which can forecast the production data [8]. Data-driven surrogate modeling of the reservoir numerical simulation process based on deep learning methods is an effective solution to the time-consuming and costly problems caused by repeated calls to the reservoir numerical simulator in the automatic history matching process [9–11]. The LSTM can predict time series data. The attention mechanism is capable of better capturing local temporal features of the production time series [12, 13]. In this work, the attention mechanism is introduced on the composition of deep CNN and LSTM. By establishing an attention-based temporal and spatial convolution recursive neural network surrogate model, the spatial characteristics of reservoir static parameters can be obtained which makes the reservoir static parameters can be mapped to the reservoir production curve. The image-sequence mapping relationship of reservoir production curve prediction is used to predict the reservoir production curve.

## 2 Methodology

With the introduction of image-to-sequence regression and the attention mechanism, an attention-based temporal and spatial convolution recursive neural network for surrogate modeling is built to achieve fast and accurate prediction of reservoir production.

The workflow for production curve forecasting using a surrogate model is shown below: Firstly, the data sample required for model is generated. Secondly, an attention-based temporal and spatial convolution recursive neural network which is built with dense CNN and multi-layer LSTM is used to capture the complex nonlinear mapping relationship between geological oil reservoir model parameters and oil reservoir production curves. In the surrogate modeling process, the accuracy of the agent model is improved by the introduction of the attention mechanism. Finally, the predictive performance of the surrogate model is tested.

## 2.1 Surrogate Model Dataset

The two-dimension reservoir stochastic permeability field used in this paper was produced by the Stanford open-source software SGeMs and the size was set to 60 × 60 × 1 grid blocks. The reservoir model assumes that the discrete permeability field of non-homogeneous reservoir follows log-normal distribution, with a logarithmic mean of 6 and a variance of 1 for the permeability.

**Table 1.** Parameters used to produce phase percolation curves.

| Range | Parameters | | | | |
|---|---|---|---|---|---|
| | $n_o$ | $n_w$ | $R_{(0,1)}$ | $(k_{ro})s_{wc}$ | $(k_{rw})s_{orw}$ |
| Upper | 7.75 | 7.75 | 1 | / | / |
| Lower | 1 | 1 | 0 | / | / |
| Value | / | / | / | 0.85 | 0.60 |

**Table 2.** Parameters used for reservoirs numerical simulation.

| Parameters | Value | Parameters | Value |
|---|---|---|---|
| Injection well downhole pressure | 430 *bar* | Water viscosity | 1 *cp* |
| Irreducible water saturation | 0.17 | Porosity | 0.2 |
| Water injection rate | 520 m$^3$/d | Residual oil saturation | 0.25 |
| Dimensions | 480 × 480 × 4 m | Production wells downhole pressure | 395 *bar* |
| The oil viscosity | 5 *cp* | | |

The relative permeability data is generated based on a Corey model with the parameters set as shown in Table 1. The well-distribution method is an inverse five-point injection mode. For each set of input data, a numerical simulator was run to generate the corresponding 750 days of constant pressure production data in conjunction with the corresponding production regime in Table 2, which was divided evenly into 25-time steps. Finally, the 2000 samples pool was split into a training set and a test set in a ratio of 3:1.

## 2.2 Network Architecture

In contrast to the traditional image-to-image regression prediction of permeability field to pressure field and saturation field [14], the output of the image-to-series regression prediction model of permeability field to production data is no longer a single image and no longer uses the pixel values on the image to characterize the desired output, but directly outputs a time series, the production data curve, with a more reasonable consideration of temporal information [15]. In addition, the image-to-image regression gets the saturation and pressure data, which requires a further calculation to obtain the production data observed on site, such as through the Peaceman equation calculation [11], which will add some additional calculations cost.

Reservoir production prediction required in the history matching process is a typical image-to-sequence regression problem. The model inputs the permeability field image and relative permeability curve, and the model directly outputs the predicted value of oil production and water production. Convolutional networks are good at extracting image features, while the LSTM has advantages over other networks in processing time-series data. Therefore, this paper focuses on combining densely connected CNN with LSTM to build a surrogate model for oil reservoir production prediction, and the network incorporates an attention mechanism to improve the accuracy.

The map between inputs and outputs of model is shown in Eq. (1):

$$f : R^{H \times W \times D \times N_f} \rightarrow R^{T \times N_d} \tag{1}$$

In the above equation, $f$ is the relationship between model input and output data; $H \times W \times D$ is the size of oil reservoir model, which is $60 \times 60 \times 1$; $N_f$ is the number of uncertain parameter fields, which is 2; $T$ is time steps, which is 25; $N_d$ is the number of the production, which is 2.

A surrogate model based on a spatial-temporal convolutional RNN is constructed by using a densely CNN and LSTM neural network combined with an attention mechanism. First, a densely connected CNN extracts spatial features of high-dimensional uncertain parameters. Then, the spatial feature maps are transformed into vector features and concatenated to relative permeability vectors. Finally, a multilayer LSTM neural network incorporating an attention mechanism is used for time series regression and prediction of the simulated data. The model is an end-to-end training model, i.e. after inputting geological parameters, the network can directly output reservoir production data without going through the next calculation step as shown in Fig. 1.

The input data of the model are the permeability field distribution and the 6 parameters of the relative permeability curves, with permeability field characterized as an image, the permeability values as image pixel values, and the relative permeability profile characterized as a vector of size 6. The permeability field is first fed into the encoding network for feature extraction and a stepwise reduction in the feature map size, with the extracted feature size being $32 \times 32$, which is then compressed into a vector of size 1024 as shown in Fig. 2. For the relative permeability, 2 simplified FNN is used first

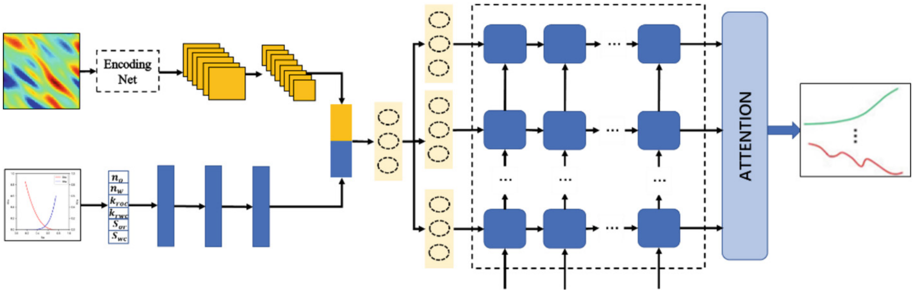


**Fig. 1.** The structure diagram of attention-based temporal and spatial convolution recursive neural network

to process the 6 parameters of the relative permeability curves which are presented as a vector. After the relative permeability vector is processed by the FNN, the feature of last FNN is a vector with a size of 1024. The features extracted from the permeability and the relative permeability are connected, and feature connection is realized by the concatenate function.

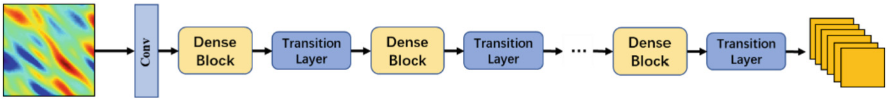


**Fig. 2.** The structure of the encoding

Once the features were extracted and connected, a fully connected neural network was first connected to change the dimension to 100, which in turn was repeated 25 times using the RepeatVector function. The feature maps were fed into a stacked multilayer LSTM neural network, which was recovered as the reservoir production over 25-time steps. The first LSTM layer received the input from the convolutional neural network features. The next layer was used to receive the hidden states from the previous layer. Multi-layer LSTM can handle sequential data with stronger non-linearities than single-layer LSTM models. Two to four layers are usually recommended, and two layers were used in this paper. The LSTM incorporated an attention mechanism to calculate the output. The size of the output is 25 × 8, 25 means 25-time steps, and 8 means the oil production and water production of 4 production wells.

Figure 3 shows the structure of an LSTM. The cell state ($c_t$) is the key to the LSTM and is used to store the state information of the current LSTM and pass it to the LSTM at the next moment. The current LSTM receives the cell state ($c_{t-1}$) from the previous moment and works together with the signal input received by the current LSTM to generate the cell state of the current LSTM.

Attentional mechanisms originated from the study of human vision, which is a brain signal processing mechanism unique to human vision. Human vision first quickly scans the global image, then acquires the target area to focus on, which is the focus of attention, and then logically devotes more attention to this area to obtain more details about the

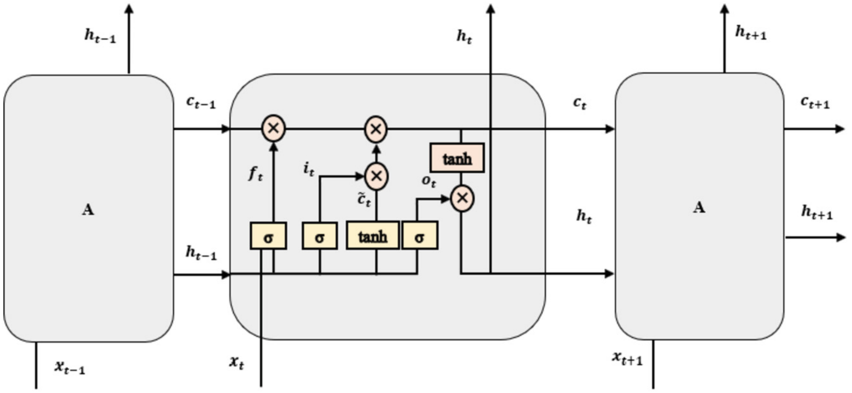


**Fig. 3.** The structure diagram of LSTM

target to be focused on, while ignoring other irrelevant information. These mechanisms are often referred to as attentional mechanisms.

The core operation of the attention mechanism is a series of weight parameters that are learned from the sequence to determine the importance of each element, and then the elements are combined by importance. Essentially the attention mechanism is a weighted summation of the values of the elements. The essential idea can be rewritten as the formula:

$$Attention(Query, Source) = \sum_{i=1}^{L_x} Similarity\left(Query^{Key_i}\right) \times Value_i \tag{2}$$

As for the specific calculation process of the Attention mechanism, if abstracted from most current methods, it can be summarized into two processes: the first is the calculation of the weighting factor based on the *Query* and *Key*, and the second is the weighted summation of the *Value* based on the weighting factor. In turn, the first process can be subdivided into two stages: the first stage calculates the similarity or correlation between *Query* and *Key*; the second stage normalizes the raw scores from the first stage. An attention layer was added to the model of LSTM to incorporate the attention mechanism, and the model principle is shown in Fig. 1.

### 2.3 Model Training and Performance Metrics

The absolute permeability values, the two shape factors characterizing the relative permeability, and the production data calculated by the oil reservoir numerical simulator were normalized to a maximum and minimum value, and the normalization method is shown below:

$$\tilde{T} = \frac{T - T_{\min}}{T_{\max} - T_{\min}} \tag{3}$$

In this equation, $\tilde{T}$ is the parameter to be normalized; $T$ is the true value of the parameter; $T_{max}$ is the maximum of the true values; $T_{min}$ is the minimum of the true values.

The surrogate model used the $L_1$ loss to measure the absolute error as shown in formula (4):

$$L_1 = \sum_{i=1}^{n} \left| y_i - \hat{y}_i \right| \tag{4}$$

In this equation, $n$ is the number of samples; $y_i$ is the true value of the $i - th$ sample; $\hat{y}_i$ is the model prediction of the $i{-}th$ sample.

A variant of stochastic gradient descent with adaptive moment estimation (*Adam*) was used as the optimizer. The batch size was set to 16 and the loss rate was set to 0.1.

Dynamic learning was used, with initial learning set to 0.001 and divided by 2 when the error was stable. Absolute error was used to iteratively update the weights of the network model using back propagation and gradient descent algorithms, until the iterative updating of the weights stopped when reached a pre-determined number of training sessions. The model and training process were implemented using Keras and TensorFlow.

The surrogate model was trained 200 times on an NVIDIA K80 GPU (4992 cores, 128G graphics) and took approximately 22 min to train.

RMSE is a measure of the $L_2$ parametric distance between the true and predicted values as shown in Eq. (5):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left\| y_i - \hat{y}_i \right\|_2^2} \tag{5}$$

The coefficient of determination ($R^2$) is generally used to evaluate the correlation between the predicted and true values. A higher score indicates a higher similarity, which is calculated as follows:

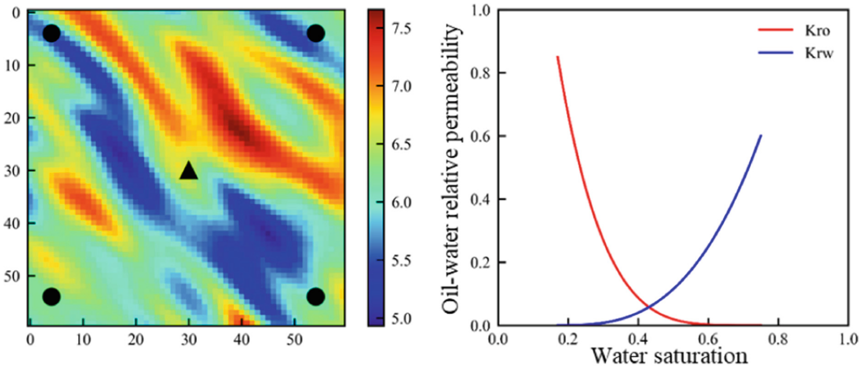
$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n-1} \left( y_i - \hat{y}_i \right)^2}{\sum_{i=0}^{n-1} (y_i - \overline{y})^2} \tag{6}$$

In this equation, $\overline{y}$ is the average target value;

## 3 Results and Discussion

The examples shown in the results are randomly selected samples from the sample library test set. The input permeability samples and the relative permeability curve corresponding to model prediction results shown in this paper are plotted as shown in Fig. 4. In the permeability field diagram, black triangles indicate injection wells and black circles indicate production wells.

The test dataset was used to evaluate the trained convolutional recursive data-driven model incorporating an attention mechanism. The proxy model predicted oil and water production over 25-time steps.



（a）Permx　（b）Relative permeability curves

**Fig. 4.** The true log-permeability and relative permeability curve

Figure 5 shows a plot of comparison of reservoir oil and water production prediction. The first row plots the oil rate of the four production wells and the second row plots the water production of the four producing wells. The solid blue line is the predicted value from the recursive convolutional network data-driven model. The red dashed line is the calculated value from the reservoir numerical simulator. It was hoped to train a surrogate model whose output would be highly similar to that calculated by the numerical simulator, so the numerical simulator's calculations are treated as real values. The closer the two curves are, the smaller the error between the data-driven surrogate model and the data predicted by the real model. Case test results show that the surrogate model can predict reservoir yields more accurately.

Figure 6 shows the loss plots for the full sample in training and testing. In this paper, both $L_1$ and RMSE errors are used to evaluate the model training results. The left curve shows the $L_1$ loss for the full sample, and the right curve shows the RMSE loss for the full sample. The solid blue line shows the curve of training set error, and the solid red line shows the loss curve of the test set with increasing number of iterations.

Figure 7 shows the coefficient of determination $\left(R^2\right)$ and learning rate for full sample in training and testing. The left plot shows $R^2$ and the right curve shows the learning rate. The solid blue line in the left curve shows the curve of $R^2$ of the training set, while the solid red line shows the curve of $R^2$ of the test set. The results showed that the coefficient of determination of the training set gradually increased to close to 1 with the training, while $R^2$ of the test set fluctuated at first and gradually increased to a steady level with the increase of iterations. The final training and test set coefficients are 0.9951 and 0.9095 respectively, and the final test $(R^2)$ is above 0.90, which means that the network has good learning ability in this case. The learning rate is set in such a way that it keeps decreasing as the error gradually smoothed out, with each change shrinking to 1/2 of the original. The graph on the right shows the curve of the learning rate, and the final learning rate is $1 \times 10^4$.

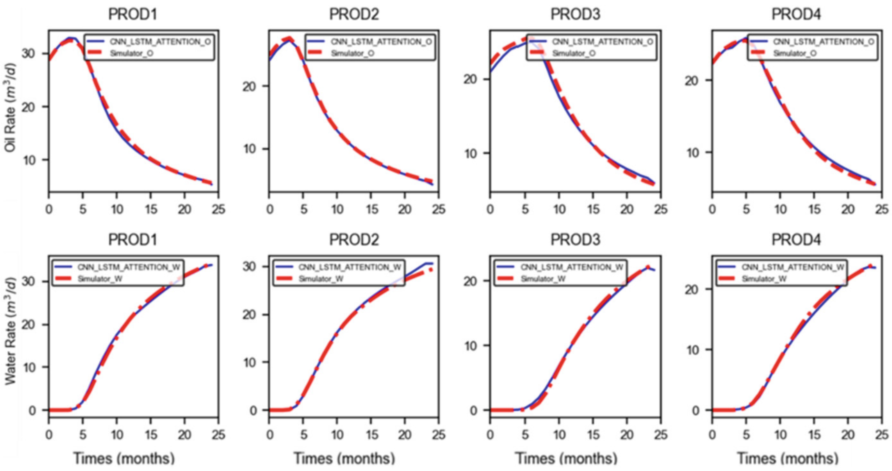
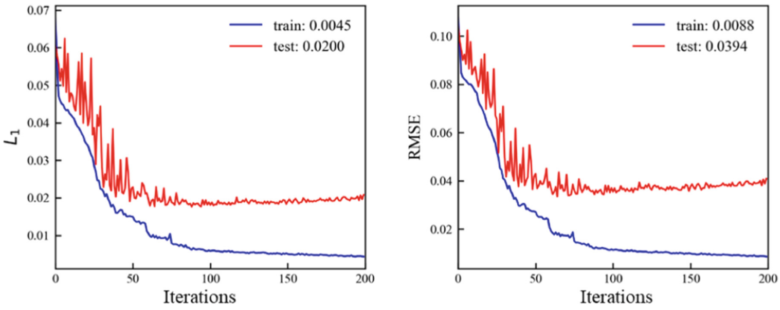


**Fig. 5.** Comparison of well rates



（a）$L_1$ （b）RMSE

**Fig. 6.** Loss curve

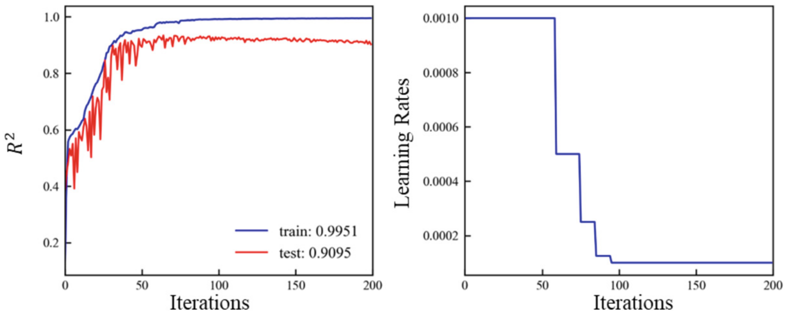


**Fig. 7.** Coefficient of determination and learning rate curve

## 4 Conclusion

This paper constructs an attention-based temporal and spatial convolution recursive neural network for surrogate modeling of the production curve prediction that can end-to-end predict production data directly from oil reservoir spatial parameters. In concrete terms, it is a surrogate modelling framework based on image-to-sequence regression, constructed based on CNN and LSTM model, combined with an attention mechanism. Instance testing of the two-dimensional oil reservoir model shows that the proposed data-driven surrogate model of spatial-temporal convolutional recurrent neural networks can predict reservoir production quickly and efficiently.

**Acknowledgments.** This work is supported by the National Natural Science Foundation of China under Grant 52274057, 52074340 and 51874335.

## References

1. Zhang, K., Zhao, X., Zhang, L., et al.: Current status and prospect for the research and application of big data and intelligent optimization methods in oilfield development. J. China Univ. Petrol. (Ed. Nat. Sci.) **44**(4), 28–38 (2020)
2. Chen, C., Gao, G., Gelderblom, P., et al.: Integration of cumulative-distribution-function mapping with principal-component analysis for the history matching of channelized reservoirs. SPE Reservoir Eval. Eng. **19**(02), 278–293 (2016)
3. Emerick, A.A.: Investigation on principal component analysis parameterizations for history matching channelized facies models with ensemble-based data assimilation. Math. Geosci. **49**(1), 85–120 (2017)
4. Oliver, D.S., Chen, Y.: Recent progress on reservoir history matching: a review. Comput. Geosci. **15**, 185–221 (2011)
5. Jung, H., Jo, H., Kim, S., et al.: Recursive update of channel information for reliable history matching of channel reservoirs using EnKF with DCT. J. Petrol. Sci. Eng. **154**, 19–37 (2017)
6. Canchumuni, S.W., Emerick, A.A., Pacheco, M.A.C.: Towards a robust parameterization for conditioning facies models using deep variational autoencoders and ensemble smoother. Comput. Geosci. **128**, 87–102 (2019)
7. Canchumuni, S.W., Emerick, A.A., Pacheco, M.A.C.: History matching geological facies models based on ensemble smoother and deep generative models. J. Petrol. Sci. Eng. **177**, 941–958 (2019)
8. Yosinski, J., Clune, J., Bengio, Y., et al.: How transferable are features in deep neural networks?. In: Advances in Neural Information Processing Systems, vol. 27(2014)
9. Zhu, Y., Zabaras, N.: Bayesian deep convolutional encoder–decoder networks for surrogate modeling and uncertainty quantification. J. Comput. Phys. **366**, 415–447 (2018)
10. Ma, X., Zhang, K., Zhao, H., et al.: A vector-to-sequence based multilayer recurrent network surrogate model for history matching of large-scale reservoir. J. Petrol. Sci. Eng. **214**, 110548 (2022)
11. Zhang, K., Wang, X., Ma, X., et al.: The prediction of reservoir production based proxy model considering spatial data and vector data. J. Petrol. Sci. Eng. **208**, 109694 (2022)
12. Vaswani, A., Shazeer, N., Pparmar, N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30(2017)

13. Ma, X., Zhang, K., Zhang, J., et al.: A novel hybrid recurrent convolutional network for surrogate modeling of history matching and uncertainty quantification. J. Petrol. Sci. Eng. **210**, 110109 (2022)
14. Li, J., Zhang, D., He, T., et al.: Uncertainty quantification of two-phase flow in porous media via the Coupled-TgNN surrogate model. Geoenergy Sci. Eng. **221**, 211368 (2023)
15. Ma, X., Zhang, K., Wang, J., et al.: An efficient spatial-temporal convolution recurrent neural network surrogate model for history matching. SPE J. **27**(02), 1160–1175 (2022)

# Production Optimization of Chemical Flooding Based on Reservoir Engineering Method

Zhi-bin An[1], Kang Zhou[2], Jian Hou[1](✉), and De-jun Wu[1]

[1] School of Petroleum Engineering, China University of Petroleum (East China),
Qingdao, China
sllx_upc@126.com

[2] College of Energy and Mining Engineering, Shandong University of Science and Technology,
Qingdao, China

**Abstract.** Production optimization is an important way to improve technical and economic benefits in the process of reservoir development. Generally, most production optimization problems of chemical flooding are solved separately using mathematical algorithms, which limits optimization efficiency. This paper introduces the prior scheme obtained from reservoir engineering method into the optimization mathematical model to improve the efficiency of production optimization problems of chemical flooding. Firstly, the reservoir numerical simulation model and optimization mathematical model for chemical flooding are established. Secondly, the injection and production allocations are carried out through statistical analysis of the present development performance of reservoir, and a prior scheme based on reservoir engineering method is obtained. Finally, the prior scheme is used as the initial scheme for optimization. The optimization mathematical model takes net present value as the objective function, and the injection-production volume and chemical agent concentration as the optimization variables. The solving algorithm adopts particle swarm optimization. It can be seen from the results that the net present value of the uniform scheme is $0.761 \times 10^8$ RMB while $0.963 \times 10^8$ RMB for the prior scheme, which has an increase of 26.54%. Moreover, the conventional method converges to $1.317 \times 10^8$ RMB after 22 iterations, while the proposed method converges to $1.328 \times 10^8$ RMB after 11 iterations. The proposed method reduces calculation amount by 50% with satisfactory accuracy. Therefore,

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 516–525, 2024.
https://doi.org/10.1007/978-981-97-0272-5_44

the proposed method using the prior scheme obtained from reservoir engineering method as the initial scheme achieves better optimization performance than conventional method. This method achieves the combination of mathematical theory and engineering experience, and providing an effective way to reduce calculation costs and increase efficiency for solving reservoir optimization production problems.

**Keywords:** Chemical flooding · Production optimization · Reservoir engineering method · Net present value

# 1 Introduction

With the implementation of polymer flooding and polymer-surfactant flooding pilot tests, chemical flooding achieves industrialization in major oilfields in China, and is currently an important method for enhanced oil recovery [1–6]. However, due to the heterogeneity of formation and the complex distribution of remaining oil after water flooding, it is necessary to further enhance oil recovery through reasonable injection-production allocation in chemical flooding stage. At present, there are two main methods for injection-production allocation. The first method is to use reservoir engineering method to adjust development scheme based on geological characteristics. The reservoir engineering method refers to the comprehensive use of various production and logging data to study the laws of various indicators in the process of development, which are used to adjust and improve the development scheme [7]. This kind of method is simple and practical, and can achieve the goal of enhanced oil recovery. However, it is usually not only difficult to obtain accurate data for complex reservoir, but also achieve best development results [8]. For these reasons, the application of this method is limited. The second method is to solve production optimization problems through algorithms. The use of optimization algorithms for production optimization problems has gradually emerged in recent years. This kind of method establishes mathematical models for optimization problems and select appropriate algorithms for solution [9–12]. Compared to reservoir engineering method, it can effectively improve the development performance without any production or logging data. However, it usually requires a large amount of calculation costs, which is difficult to apply in practical production in commercial oilfields. Therefore, this paper combines reservoir engineering method and optimization algorithms to solve production optimization problems of chemical flooding. The prior scheme is obtained through statistical analysis of current development performance of reservoir in the stage of chemical flooding, which is used as the initial scheme for optimization to achieve better performance.

# 2 Method

## 2.1 Reservoir Engineering Method

In the process of oilfield development, the heterogeneity of the formation leads to differences in the development performance of different parts of reservoir. It is an effective way to improve oil recovery by formulating development scheme based on the development performance. The injection-production allocation scheme can be obtained through reservoir engineering method to statistically analyze the development performance in the stage of chemical flooding, which is taken as the prior scheme. Compared to uniform scheme, the prior scheme reflects the present development performance of reservoir and has better technical and economic benefits. The specific steps are as follows:

Firstly, the well pattern is divided based on the location and corresponding relationship between injection and production wells. Generally, the area well pattern is adopted considering the complexity of the actual oilfields well pattern. In each area well pattern, the non-central wells are connected sequentially, and the midpoint of the connection lines of non-central wells are connected with the central well. The divisions of well pattern and regions within each well pattern are achieved through this method. Secondly, the parameter values of each grid in the numerical simulation model required for injection-production allocation are counted in the chemical flooding stage. And the sum of parameter values in each region are calculated based on the grids contained within each region. Generally, production allocation parameters include remaining oil saturation, remaining geological reserves, formation pressure and others. The injection allocation parameters include pore volume, formation pressure, formation coefficient and others. Finally, each non-central well is allocated injection or production proportionally based on the parameter values of corresponding region under the condition of constant total injection and production allocation, and each central well is allocated proportionally based on the sum of parameter values of all regions in the corresponding well pattern. Taking injection well as the central well and production well as non-central well as an example, the calculation formula of injection and production allocation is shown as:

$$Q_{m,k}^{p} = \frac{Para_{m,k}}{\sum_{m=1}^{M} \sum_{k=1}^{K_m} Para_{m,k}} Q_{total}^{p} \tag{1}$$

$$Q_{m}^{i} = \frac{\sum_{k=1}^{K_m} Para_{m,k}}{\sum_{m=1}^{M} \sum_{k=1}^{K_m} Para_{m,k}} Q_{total}^{i} \tag{2}$$

where $M$ is the number of well patterns; $K_m$ is the number of regions in each well pattern; $Q_{m,k}^{p}$ is the production allocation of the production well of region $k$ in well pattern $m$; $Q_{m}^{i}$ is the injection allocation of the injection well in well pattern $m$; $Q_{total}^{i}$ and $Q_{total}^{p}$ are the total injection and production amount of the model, respectively; $Para_{m,k}$ is the parameter value of region $k$ in well pattern $m$.

## 2.2 Optimization Mathematical Model

**Model Establishment.** Generally, the evaluation of development performance includes technical and economic. Net present value (NPV) as an economic indicator is a common method to evaluate the development performance for optimization problems [13–15]. NPV refers to the cumulative net cash flow discounted from the net cash flow of each year to the starting year within the project implementation cycle. The calculation process conforms to the benchmark rate of return or discount rate of the industry. Taking polymer flooding as an example, the injection-production optimization mathematical model can be established.

$$NPV = \sum_{i=1}^{t} \left\{ Q_o P_o \alpha - \left[ Q_o C_m + Q_p P_p + (R_s + R_t P_o) Q_o \alpha \right] \right\} \times (1 + r)^{-i} - nI_s \quad (3)$$

where $t$ is the development time after polymer flooding; $P_o$ and $P_p$ are the prices of oil and polymer, respectively; $Q_o$ is the annual oil increase of polymer flooding relative to water flooding; $Q_p$ is the annual injection volume of polymer; $C_m$ is the incremental operating cost of oil; $\alpha$ is the oil commodity rate; $R_s$ is the resource tax; $R_t$ is the comprehensive tax rate; $r$ is the benchmark rate of return; $n$ is the total number of wells for polymer flooding; $n$ is the incremental investment cost for a single well.

Furthermore, the range of injection and production parameters are limited by equipment, technology, and other factors. That is to say, each optimization variable must comply with bound constraints.

$$u_k^{low} \leq u_k \leq u_k^{up},\ k = 1, \cdots, N_u \quad (4)$$

where $u_k$ is the optimization variable; $u_k^{low}$ is the lower bound of $u_k$; $u_k^{up}$ is the upper bound of $u_k$; $N_u$ is the number of optimization variables.

**Solving Method.** Considering the complexity of chemical flooding optimization problems, particle swarm optimization (PSO) is adopted in this paper. PSO is a gradient free optimization algorithm proposed by Eberhart and Kennedy, which has strong global search ability [16–19]. It does not rely on the gradient information, and only seeks the optimal solution through information sharing among individuals in the group. In the production optimization problem, each particle represents a feasible combination of injection and production parameters. The formula of PSO is as follows:

$$u_m^{l+1} = u_m^l + v_m^{l+1} \quad (5)$$

$$v_m^{l+1} = \omega v_m^l + c_c r_c^l (u_{mpb}^l - u_m^l) + c_s r_s^l (u_{gpb}^l - u_m^l) \quad (6)$$

where $u_m^l$ and $u_m^{l+1}$ are the position of particle $m$ after $l$ and $l+1$ iteration, respectively; $v_m^{l+1}$ is the velocity of particle $m$ in $l+1$ iteration; $\omega$ is the inertia weight; $c_c$ and $c_s$ are acceleration constants tend to itself optimal of global optimal, respectively; $r_c^l$ and $r_s^l$ are random numbers vary from 0 to 1; $u_{mpb}^l$ is the best position searched by particle $m$ after $l$ iteration; $u_{gpb}^l$ is the best position searched by all particles after $l$ iteration, which represents the current optimal value of the objection function.

## 3 Results

The numerical simulation model for polymer flooding is established based on A oilfield, as shown in Fig. 1. The model adopts corner grid system with a total of 488160 grids. Among them, there are 120, 113, and 36 grids in the X, Y, and Z directions, respectively.

The model includes 12 injection wells and 21 production wells. It is worth noting that there are 9 injection wells implement layered injection process, which is represented by well name followed by "N". The basic physical parameters are shown in Table 1. The model adopts water flooding firstly, and turn to polymer flooding when the water cut arrives at 95%. After 0.3PV of polymer flooding, it turns to water flooding and stops until the water cut arrives at 98%. Therefore, the number of optimization variables in production optimization problems is 63. Among them, since there are 9 layered injection wells, the number of production rate, injection rate, and polymer concentration are 21, 21, and 21, respectively.

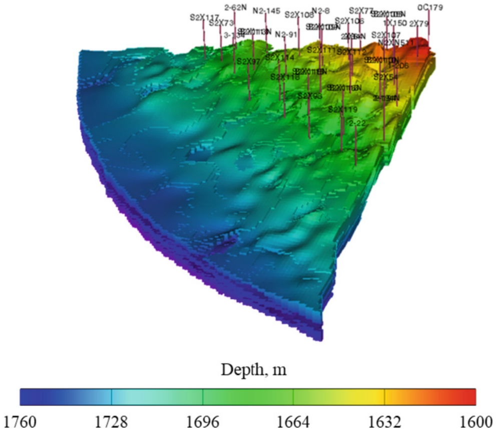


**Fig. 1.** Numerical simulation model

**Table 1.** Physical parameters of reservoir

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| Geological reserves/$10^4 m^3$ | 877 | Permeability/$10^{-3} \mu m^2$ | 1719 |
| Pore volume/$10^4 m^3$ | 3450 | Oil viscosity/mPa·s | 40 |
| Oil saturation/% | 65 | Oil density/kg·$m^{-3}$ | 950 |
| Porosity/% | 21 | Pressure/MPa | 16.7 |

The prior scheme is designed for polymer flooding based on reservoir engineering method. Furthermore, a uniform scheme is designed for comparison to illustrate the performance of the proposed method, which injection volume, production volume, and polymer concentration are evenly distributed in each well. In uniform scheme, the injection volume is split according to the perforating layers. In this paper, the inertia weight and acceleration coefficients are 0.7298 and 1.4296, respectively [20]. The value of economic parameters in NPV is shown in Table 2.

The comparison of iteration curves of using the uniform scheme and prior scheme as the initial scheme are shown in Fig. 2. The NPV of the uniform scheme is 0.761

**Table 2.** Value of economic parameters in NPV

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| Benchmark rate of return/% | 12 | Oil price/(RMB·$t^{-1}$) | 2500 |
| Incremental operating cost of oil/(RMB·$t^{-1}$) | 600 | Resource tax/(RMB·$t^{-1}$) | 14 |
| Comprehensive tax rate/% | 14 | Oil commodity rate/% | 97 |
| Incremental investment cost/(RMB·$well^{-1}$) | 1000000 | Polymer price/(RMB·$t^{-1}$) | 20000 |

× $10^8$ RMB, and converges to 1.317 × $10^8$ RMB after 22 iterations. The NPV of the prior scheme is 0.963 × $10^8$ RMB, and converges to 1.328 × $10^8$ RMB after 11 iterations. It can be seen that the NPV of prior scheme increases by 26.54% compared to the uniform scheme, and reduces calculation amount by 50% after convergence with satisfactory accuracy. Therefore, the proposed method of using prior scheme obtained from reservoir engineering method as the initial scheme enables optimization problems to be carried out at a better initial level, which can significantly reduce calculation costs.

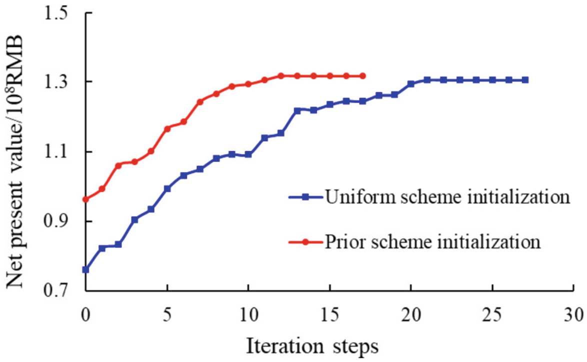
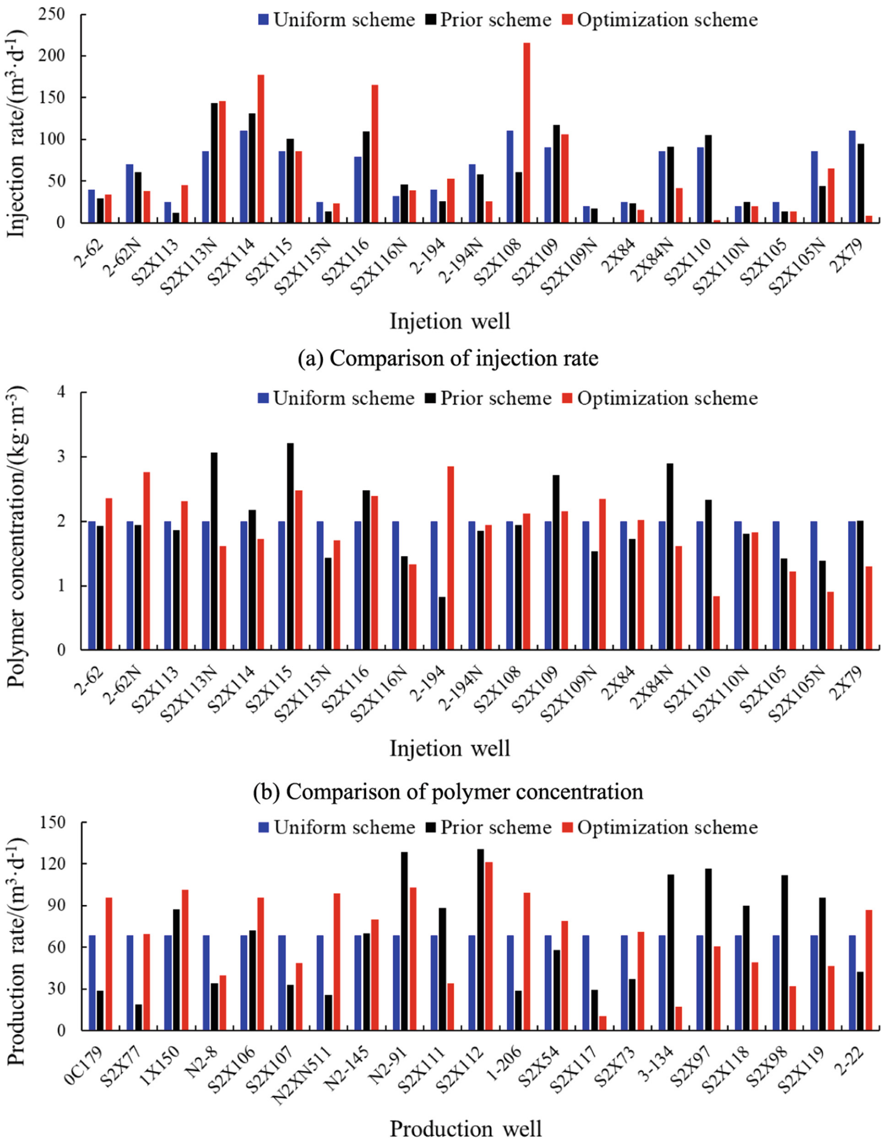


**Fig. 2.** Comparison of iteration curves

The results of prior scheme initialization are taken as the optimization scheme, and compare with the uniform scheme and prior scheme. The comparisons of injection rate, production rate, and polymer concentration for each well are shown in Fig. 3. It can be seen that the variation trend of the prior scheme and optimization scheme is roughly the same compared to uniform scheme, reflecting the reservoir experience of injection-production adjustment based on geological conditions and development performance of reservoir.

Figure 4 is the comparison of recovery factor and water cut of the uniform scheme, prior scheme, and optimization scheme. It can be seen that the recovery factor of the uniform scheme is 26.17%, and the lowest value of water cut is 88.21%. The recovery factor of the prior scheme and optimization scheme are 26.70% and 28.53%, increasing



(a) Comparison of injection rate

(b) Comparison of polymer concentration

(c) Comparison of production rate

**Fig. 3.** Comparisons of optimization variables

by 0.53% and 2.36% compared to the uniform scheme. The lowest value of water cut of the prior scheme and optimization scheme are 86.87% and 85.87%, reducing by 1.34% and 2.34% compared to the uniform scheme. Furthermore, the increase speed of water cut of the prior scheme and optimization scheme is slower than that of uniform scheme. Figure 5 compares the NPV of the uniform scheme, prior scheme, and optimization scheme. The prior scheme and optimization scheme have larger NPV and shorter payback period. It is worth noting that there is a slight decrease in NPV at the end of development, which is because the annual oil increase of polymer flooding relative to water flooding is negative. Briefly, the prior scheme can achieve better development performance both in technical and economic compared to uniform scheme, thereby achieving the goal of improving optimization efficiency.

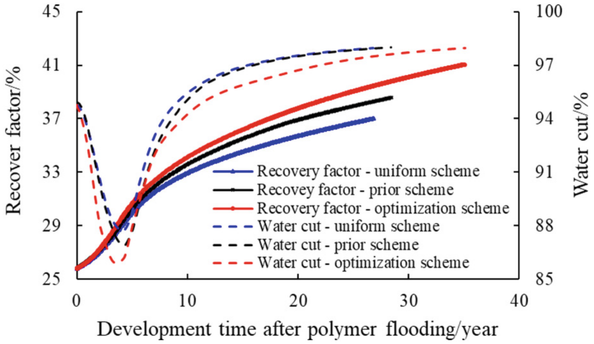


**Fig. 4.** Comparison of recovery factor and water cut

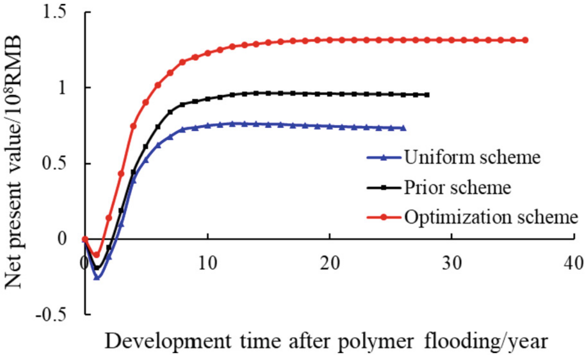


**Fig. 5.** Comparison of NPV

## 4 Conclusions

This paper introduces reservoir engineering method into optimization algorithms to solve production optimization problems of chemical flooding. The optimization process can be effectively accelerated by using the prior scheme obtained from reservoir engineering method as the initial scheme. The results show that the NPV of the prior scheme is 0.963 $\times$ $10^8$ RMB, which is 26.54% higher than that of the uniform scheme. Meanwhile, the method of uniform scheme initialization converges to 1.317 $\times$ $10^8$ RMB after 22 iterations, while prior scheme initialization converges to 1.328 $\times$ $10^8$ RMB after 11 iterations. The calculation amount reduces by 50% with satisfactory accuracy. The proposed method combines mathematical theory with engineering experience, providing an effective way to reduce costs and increase efficiency for solving production optimization problems of chemical flooding.

**Acknowledgments.** The authors greatly appreciate the financial support of the National Natural Science Foundation of China (Grant No. 52104027), the Project supported by the Joint Funds of the National Natural Science Foundation of China (Grant No. U21B2070) and the Shandong Provincial Natural Science Foundation (Grant No. ZR2021ME072).

## References

1. Cao, X.: Design and performance evaluation on the heterogeneous combination flooding system. Acta Pet. Sin. **29**, 115–121 (2013)
2. Wu, D., Zhou, K., Hou, J., et al.: Experimental study on combining heterogeneous phase composite flooding and streamline adjustment to improve oil recovery in heterogeneous reservoirs. J. Pet. Sci. Eng. **194**, 107478 (2020)
3. Wu, D., Zhou, K., Zhao, F., et al.: Determination of permeability contrast limits for applying polymer solutions and viscoelastic particle suspensions in heterogeneous reservoirs. Energy Fuels **36**(14), 7495–7506 (2022)
4. Seright, R., Brattekas, B.: Water shutoff and conformance improvement: an introduction. Pet. Sci. **18**, 450–478 (2021)
5. Ma, Y., Hou, J., Shang, D., et al.: Effect of the loss of viscosity and viscoelasticity on displacement efficiency in polymer flooding. Pet. Sci. Bull. **1**, 133–141 (2017)
6. Zhou, K., Wu, D., An, Z.: Experimental study on matched particle size and elastic modulus of preformed particle gel for oil reservoirs. Gels **8**(8), 506 (2022)
7. Song, Z., Li, Z., Lai, F., et al.: Derivation of water flooding characteristic curve for high water-cut oilfields. Petroleum Explor. Dev. **40**(2), 216–223 (2013)
8. Golzari, A., Sefat, M.H., Jamshidi, S.: Development of an adaptive surrogate model for production optimization. J. Pet. Sci. Eng. **133**, 677–688 (2015)
9. Zhang, F.: Optimization method of injection and production parameters for polymer/surfactant binary flooding. J. China Univ. Petrol. (Ed. Nat. Sci.) **42**(5), 98–104 (2018)
10. Janiga, D., Czarnota, R., Stopa, J., et al.: Performance of nature inspired optimization algorithms for polymer enhanced oil recovery process. J. Pet. Sci. Eng. **154**, 354–366 (2017)
11. Nasir, Y., Yu, W., Sepehrnoori, K., et al.: Hybrid derivative-free technique and effective machine learning surrogate for nonlinear constrained well placement and production optimization. J. Pet. Sci. Eng. **186**, 106726 (2020)
12. Zhao, H., Zhang, Y., Cao, L., et al.: Constrained short-term and long-term multi-objective production optimization using general stochastic approximation algorithm. Cluster Comput. **22**, 6267–6281 (2019)
13. Karambeigi, M., Zabihi, R., Hekmat, Z.: Neuro-simulation modeling of chemical flooding. J. Pet. Sci. Eng. **78**(2), 208–219 (2011)
14. Chen, C., Wang, Y., Li, G., et al.: Closed-loop reservoir management on the Brugge test case. Comput. Geosci. **14**(4), 691–703 (2010)
15. Lamas, L.F., Schiozer, D.J., Delshad, M.: Impacts of polymer properties on field indicators of reservoir development projects. J. Pet. Sci. Eng. **147**, 346–355 (2016)
16. Eberhart, R.C., Kennedy, J.: A new optimizer using particle swarm theory. In: Proceedings of the Sixth International Symposium on Micro Machine and Human Science, pp. 39–43. IEEE, Piscataway, NJ, United States (1995)

17. Hou, J., Zhou, K., Zhao, H., et al.: Hybrid optimization technique for cyclic steam stimulation by horizontals well in heavy oil reservoir. Comput. Chem. Eng. **84**, 363–370 (2016)
18. An, Z., Zhou, K., Hou, J., et al.: Research on the main controlling factors for injection and production allocation of polymer flooding. J. Energy Resour. Technol.-Trans. ASME **145**(4), 043201 (2023)
19. Brantson, E.T., Ju, B., Opoku Appau, P., et al.: Development of hybrid low salinity water polymer flooding numerical reservoir simulator and smart proxy model for chemical enhanced oil recovery (CEOR). J. Pet. Sci. Eng. **187**, 106751 (2020)
20. Gong, Y., Li, J., Zhou, Y., et al.: Genetic learning particle swarm optimization. IEEE Trans. Cybern. **46**(10), 2277–2290 (2016)

# Research on Oilfield Surface Environment Assessment Method Based on High-Precision Global Land Cover Data Analysis

Hui Wang(✉), Wen-jing Zhang, Tao Xue, Yi-nong Li, Yan-hua Guan, Ji-ying Liu, Hong-li Jiang, and Li Bao

Exploration and Development Research Institute of Daqing Oilfield Co. Ltd., Daqing, China
157143754@qq.com

**Abstract.** In recent years, the country has successively passed a series of laws and regulations on the reform of the mining rights system, especially related to exploration and construction, land use, and ecological environment restrictions, which have a profound impact on the exploration decision-making and development of oil fields. Analyzing the available dimensions of mining rights, comprehensively evaluating the ground environment, determining whether the ground conditions can be constructed, and the difficulty of construction, has become one of the important tasks in oilfield exploration deployment and mining rights evaluation. GlobeLand30 is the world's first high-resolution land cover dataset, with the advantages of comprehensive information coverage and high display accuracy. This article takes GlobeLand30 as the starting point, and through key steps such as data collection and download, coordinate projection conversion, and vectorization into maps, it achieves the accurate display of all surface elements in oilfield exploration areas for the first time, accurately implementing the distribution and morphology of various surface elements. Meanwhile, combined with the analysis of oilfield exploration and construction technology and policy restrictions, the surface types of difficult operation areas in the exploration area have been clarified, and precise evaluation of the surface environment has been achieved through hierarchical classification and vectorization mapping. The results indicate that the main types of surface cover in the Daqing Oilfield exploration area of the Songliao Basin are cultivated land, grassland, water bodies, wetland, and artificial surfaces;

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20–22 September 2023.

This paper was selected for presentation by the IFEDC Committee following review of information contained in an abstract submitted by the author(s). Contents of the paper, as presented, have not been reviewed by the IFEDC Technical Team and are subject to correction by the author(s). The material does not necessarily reflect any position of the IFEDC Technical Committee its members. Papers presented at the Conference are subject to publication review by Professional Team of IFEDC Technical Committee. Electronic reproduction, distribution, or storage of any part of this paper for commercial purposes without the written consent of IFEDC Organizing Committee is prohibited. Permission to reproduce in print is restricted to an abstract of not more than 300 words; illustrations may not be copied. The abstract must contain conspicuous acknowledgment of IFEDC. Contact email: paper@ifedc.org.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 526–537, 2024.
https://doi.org/10.1007/978-981-97-0272-5_45

The most difficult construction operations are water bodies and wetland, followed by artificial surfaces, and the construction difficulty of cultivated land, grasslands, and forest also needs to consider policy restrictions. The ground environmental assessment method based on GlobeLand30 has important scientific guidance for oilfield exploration deployment decision-making and mining protection plan formulation.

**Keywords:** High Precision Global Surface Cover Data · GlobeLand30 · Oilfield · Ground Environment Assessment · Difficult Operation Area

## 1 Introduction

Ground environment assessment is one of the important contents of oil and gas exploration evaluation. The underground geological conditions of exploration targets are the main factors determining whether oil and gas can be obtained, while the ground conditions and the difficulty of construction are the direct factors affecting exploration deployment. In recent years, the country has vigorously promoted the reform of the oil and gas mining rights system, especially policies such as competitive transfer of exploration rights, mandatory 25% reduction upon expiration, and strict adherence to the ecological red line, which have led to many unfavorable situations for exploration rights in various domestic oil fields, such as "fast reduction, difficult acquisition, and difficult construction" [1–3]. How to fully protect high-quality exploration rights and accurately determine the difficult operation areas within existing exploration rights has important practical and strategic significance for oilfield exploration deployment decision-making and long-term development.

At present, due to limited data mastery and access, there are generally problems with incomplete surface information elements, inaccurate location, and insufficient accuracy in surface environmental assessment of various oil fields [4–8]. Specifically, incomplete and inaccurate data on difficult operate areas such as rivers, lakes, and wetlands; Artificial surfaces in urban areas, airports, villages, and other areas are mostly schematic labels without complete and clear boundaries, and the dynamic evaluation of planning and construction areas is relatively lagging behind; The existing data on "policy restricted areas" such as nature protected areas and ecological redlines are limited, and the comprehensive assessment of the ground environment is not in place [9]. The above issues directly affect the selection of exploration target locations, construction methods, and plan formulation, and are also practical challenges that oilfields must attach importance to and solve after the reform of the mining rights system.

## 2 Brief Introduction to Global Surface Cover Data

Surface cover and its changes are essential basic information and key parameters for environmental change research, geographic monitoring of national conditions, and sustainable development planning. So far, the United States and the European Union have completed global surface cover data products with resolutions of 1000 m and 300 m.

China independently developed and officially released the world's first 30 m resolution global land cover dataset, GlobeLand30, in 2014, and donated it to the United Nations, becoming the first global geographic information high-tech public product provided by China to the United Nations.

The GlobeLand30 dataset is an important achievement of the China National High tech Research and Development Program's "Global Land Cover Remote Sensing Mapping and Key Technology Research" project. This project is led by the National Basic Geographic Information Center, and 18 units from 7 departments including the Surveying and Mapping Bureau, the Chinese Academy of Sciences, the Ministry of Education, the Ministry of Agriculture, and the Forestry Bureau jointly participated in the completion. The GlobeLand30 dataset covers the land area at latitude 80° north and south, representing 10 types of surface cover in different colors, including arable land, forests, grasslands, shrublands, wetlands, water bodies, tundra, artificial land, bare land, glaciers, and permanent snow cover (Table 1). It contains richer and more detailed spatial distribution information of surface cover, which can better depict human land use activities and the landscape patterns formed by them.

**Table 1.** Explanation of 10 Different Surface Types of GlobeLand30

| Code | Class | Content | Color |
|---|---|---|---|
| 10 | Cultivated Land | Lands used for agriculture, horticulture and gardens. | 🟪 |
| 20 | Forest | Lands covered with trees, with vegetation cover over 30%. | 🟩 |
| 30 | Grassland | Lands covered by natural grass with cover over 10%. | 🟩 |
| 40 | Shrubland | Lands covered with shrubs with cover over 30%. | 🟩 |
| 50 | Wetland | Lands covered with wetland plants and water bodies. | 🟦 |
| 60 | Water Bodies | Water bodies in the land area. | 🟦 |
| 70 | Tundra | Lands covered by lichen, moss, hardy perennial herb and shrubs in the polar regions. | 🟫 |
| 80 | Artificial Surfac-es | Lands modified by human activities. | 🟥 |
| 90 | Bareland | Lands with vegetation cover lower than 10%. | ⬜ |
| 100 | Permanent Snow & Ice | Lands covered by permanent snow, glacier and icecap. | ⬜ |

## 3 General Idea

A detailed ground environment assessment requires a comprehensive understanding of the actual surface coverage information of the exploration area, and precise implementation of the distribution and morphology of various difficult operation areas. After vectorizing the surface elements into maps and overlaying them with drilling, seismic, infrastructure and other data information of the oil field, it is necessary to accurately determine the types of difficult operation areas and their construction difficulties, thereby guiding exploration deployment decisions and plan formulation. The following two overall goals should be achieved: to construct comprehensive and accurate surface coverage information of exploration areas, clarify surface types, and guide exploration deployment decisions; Identify key surface elements, focus on characterizing difficult operation areas and vectorizing. Then, through classification, and statistical analysis, carry out a detailed evaluation of the surface environment, providing a scientific basis for formulating mining protection plans.

In response to the above objectives, this article first cites high-precision global surface coverage data as the basic data. Through key steps such as Data information acquisition and organization, Coordinate projection conversion and Map-making, and Vectorization of key surface elements, a complete and accurate map of surface coverage information in exploration areas is obtained. Then, combined with geographic information and satellite maps, mutual proofreading is carried out to determine the types and boundaries of difficult operation areas. Finally, a comprehensive evaluation map of the ground environment in the exploration area was established, and we can use it for achieving the goal of classifying and grading the difficult operation areas within the exploration area (Fig. 1). The key technical steps include Coordinate projection conversion and Color information vectorization into a graph.

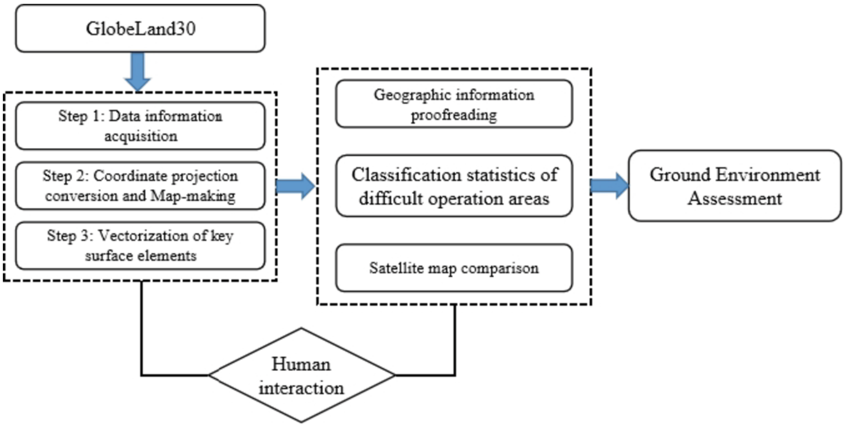

GlobeLand30
Step 1: Data information acquisition
Step 2: Coordinate projection conversion and Map-making
Step 3: Vectorization of key surface elements
Geographic information proofreading
Classification statistics of difficult operation areas
Satellite map comparison
Ground Environment Assessment
Human interaction

**Fig. 1.** Schematic diagram of technical process for ground environment assessment.

# 4 Specific Technical Solution

## 4.1 Data Information Acquisition

Log in to the website of “National Catalogue Service For Geographic Information”, register user information with real name, and indicate the data using field and purpose [10]. Wait for approval before downloading. Click on the “GlobeLand30” on the homepage to open the system interface. Click on Asia and China in the menu bar on the right, compare them with the map, select the corresponding global surface coverage data based on the location of the work area, and download individual tile data separately. It should be noted that each tile data has different geographic location parameter information. When downloading, it is necessary to record its projection type, geodetic reference, band number, and other parameters in order to correctly set and convert projection parameters.

The GlobeLand30 tile data product consists of format files such as “*.tif, *.tfw, *.shp, *.xml”. Among them, the data used in this study is “*.tif” file with its own geocoding information, which contains various types of surface coverage information and is the main image result file of the segmented data.

## 4.2 *.tif File Loading and Mapping

For ease of use and accurate location, commonly used maps in various regions have traditional geographic coordinate projection parameters. For the Songliao Basin, the projection parameters commonly used in the DoubleFoxDraw Software for oil fields are “Geodetic Datum China1954 or China1980, Projection Gauss-Kruger, Central Meridian E123, and 6 Degree Zone Number 21”. The geographic coordinate information of the downloaded GlobeLand30 tile data is “Geodetic Datum WGS84, Projection UTM, Central Meridian E123, and 6 Degree Zone Number 51”. This requires the correct loading and projection conversion of the map, which can be converted into the commonly used coordinate projection system in the oil field, in order to correctly match and use the information elements such as drilling and mining rights in the oil field.

The loading and drawing of “*.tif” files are mainly divided into two steps.

The first step is to load the original TIF image file. Decompress the segmented data compression package, load the original image file with the suffix “.tif” into the commonly used DoubleFoxDraw Software in the oil field, and correctly set the coordinate projection parameters of the original image file based on the geodetic datum, projection, central longitude, segmented band number, and other information of the segmented data. After setting up, save as “File 1”.

The second step is to perform coordinate projection conversion on “File 1”. Firstly, open the “Projection Conversion” module in the DoubleFox Cartography System, select “File 1” at the input file, and set parameter information such as ellipsoid, projection, coordinates, and numbers one by one according to their original projection parameters; Secondly, at the place of output file, create a new location and file name for saving the output file (which can be defined as “File 2”), and set the ellipsoid, projection, coordinates, and number information of “File 2” based on the commonly used map projection information in the oilfield. After setting, click “Convert”, and the converted “File 2” will be the correct matching map with the existing available map in the oilfield.

In addition, if the research area spans multiple segmented data (i.e. "cross zone"), it is necessary to download the corresponding multiple segmented data, load the TIF file and coordinate projection conversion separately, select the main-zone position map, convert the sub-zone map to the main-zone coordinate projection, and then perform multi-map stitching to finally build a complete surface coverage information map of the research area (Fig. 2).

The surface coverage data map of the Daqing exploration area in the northern part of the Songliao Basin (Fig. 2) drawn through the above steps not only includes all

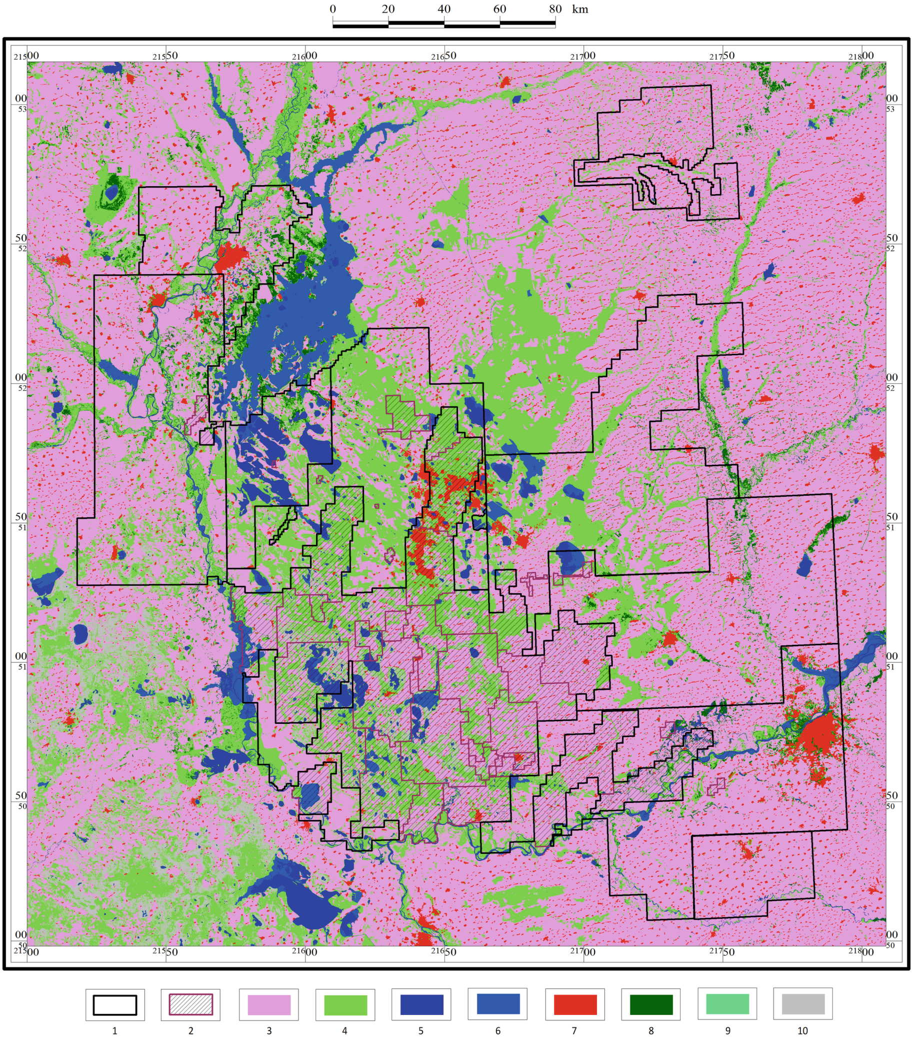


**Fig. 2.** Surface Cover Data Map of Daqing Exploration Area in the Northern Songliao Basin. Legend Description: 1-Exploration Right, 2-Mining Rights, 3-Cultivated Land, 4-Grassland, 5-Water Bodies, 6-Wetland, 7-Artificial Surfaces, 8-Forest, 9-Shrubland, 10-Bareland.

accurate surface coverage data of Cultivated Land, Grassland, Wetland, Water Bodies, etc. in the research area, but also can be used in conjunction with existing mining rights, drilling, seismic and other maps in the oil field, which can intuitively, accurately, and comprehensively display the surface information of the oilfield area for exploration and deployment personnel to use, greatly improving work efficiency.

### 4.3 Classification, Gradation, and Vectorization of Difficult Operation Areas

The above surface coverage map can well reflect the actual situation of the exploration area's surface, and have high practical value. However, considering the severe situation of current exploration and mining right management and environmental restrictions, comprehensive analysis of the availability of ground conditions in different exploration and mining rights or certain specific areas, from qualitative research to quantitative statistical analysis of surface composition, especially the classification and quantitative evaluation of difficult operation areas in all oilfields, has become a practical issue that needs to be taken seriously. The core problem is to comprehensively consider the current policies and oilfield exploration and construction capabilities, and to comprehensively evaluate the surface types and areas of the "difficult operation area".

For onshore oilfields, generally speaking, among the 10 types of surface cover, Water Bodies, Wetlands, and Artificial Surfaces are all difficult operation areas, with the greatest difficulty in exploration and construction; Cultivated Land, Forests, and Grasslands have certain policy-limited construction difficulties, which require specific analysis and comprehensive evaluation to maximize the clarity of the construction scope and difficulty level; The construction difficulty of Shrubland, Tundra, and Bareland is minimal and generally not limited by policies; Permanent Snow and Ice are currently not involved in domestic oilfields.

From the perspective of methodological research, this study identified three types of surfaces: Water Bodies, Wetland, and Artificial Surfaces as difficult operation areas, and vectorized them one by one using DoubleFoxDraw Software commonly used in Daqing oilfields.

The first step is color replacement. For example, prioritizing the selection of dark blue colors that represent Water Bodies for vectorization. In the surface coverage information map (Fig. 2), right-click on the image to enter the image processing interface, click "Color Replacement", keep the dark blue color unchanged, and replace the colors of other surface types with white one by one. Now, the image only retains two colors: dark blue and white. The second step is color binarization. Right click on the image to enter the image processing interface, select "Color Binarization", set the maximum and minimum range values for input binarization, click "OK", and wait for the software to complete running. The third step is to vectorize the image color blocks. Right click on the image to enter the image processing interface, click on "Image Semi-automatic Vectorization", and the software will start automatic vectorization of colors. After the software finish running, a complete and closed vectorized line will be formed around each dark blue color block, indicating that the color vectorization of the Water Bodies completed. Repeat the above operation and vectorize the other difficult operation areas one by one, ultimately obtaining a complete vectorized map of all difficult operation areas. Figure 3 is a vectorized map of two difficult operation areas: Water Bodies and Wetland, with dark blue representing Water Bodies and light blue representing Wetland.

Due to the objectivity and complexity of surface information, vectorization of two or more types of color blocks often involves situations such as inclusion, intersection, and overlap. For solution, manual identification and removal of overlapping elements are required. For large and complex research areas, the workload of manual identifying and removing overlapping elements is relatively high.

After measurement and comparison, using this method to vectorize the color blocks of difficult operation area in DoubleFoxDraw Software, the boundaries are clear and complete. Each difficult operation area has independent and continuous boundary lines, and the position error is less than 50 m (Fig. 4), fully achieving the purpose of classifying, grading, and statistical analysis of difficult operation areas.

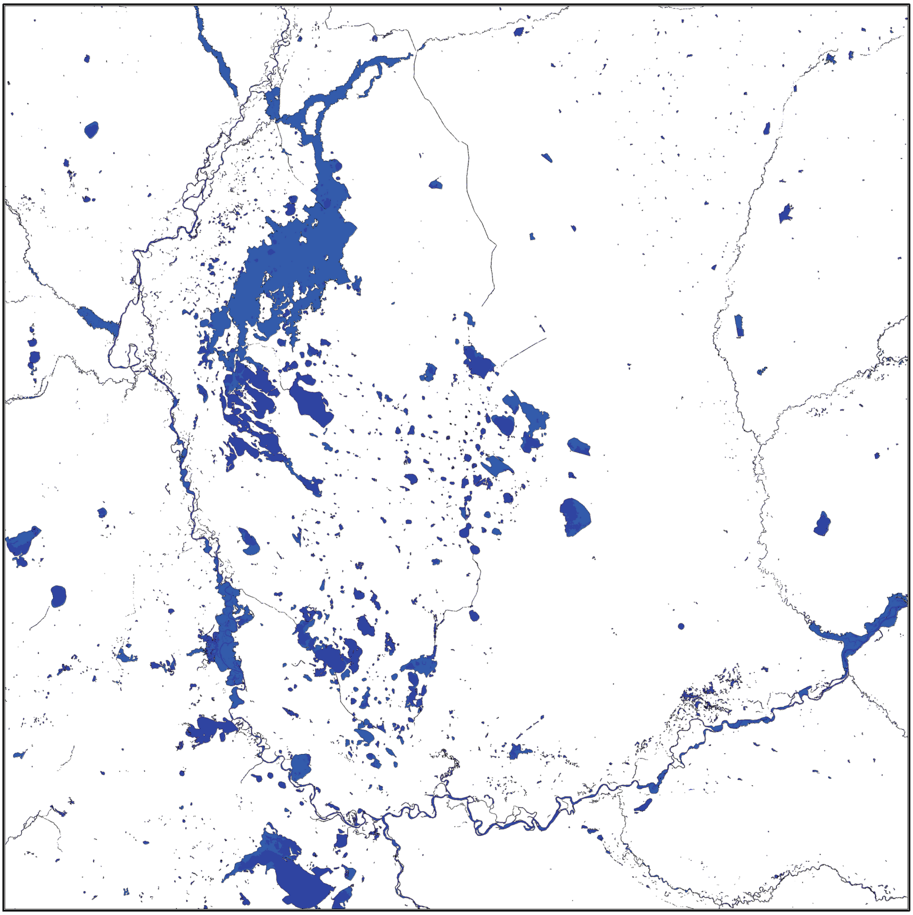


**Fig. 3.** Example of color vectorization of water bodies and wetlands in the northern Songliao Basin. Description: dark blue color representing Water Bodies, light blue color representing Wetland.

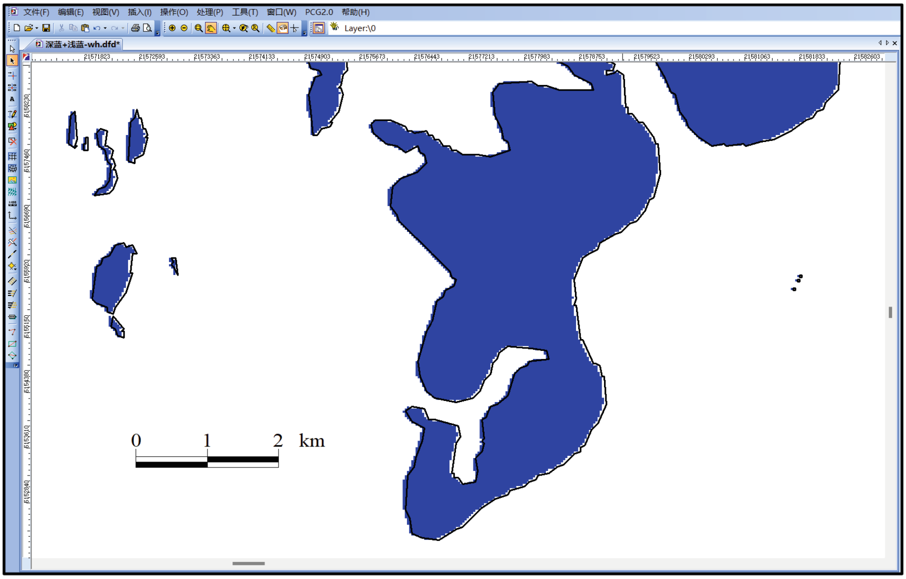


**Fig. 4.** Enlarged display of Water Bodies boundaries after vectorization. All of the boundary lines of Water Bodies are independent and continuous, and the position error is less than 50 m after measurement.

### 4.4 Preparation of Ground Environment Comprehensive Assessment Map

In fact, there are many conflicts in oil and gas exploration and development, especially the conflict of diversified mining rights, the conflict of whether the ecological environment allows construction, and so on, which directly affect the work process and interests of the oilfields [11]. For oilfields, ground environment assessment is a very important foundational work in order to avoid more conflicts. The purpose of ground environment assessment in oilfield exploration areas is to accurately discriminate the type, location, and area of difficult operation areas, and further to guide exploration and development, deployment plans, exploration and mining rights protection, and other works through classification and statistical analysis of difficult operation areas. The specific method is to establish different types of difficult operation areas in the DoubleFoxDraw Software map to achieve classification management, and then establish different levels of sub layers within the type layer based on the area size of the elements, and classify and count the difficult operation areas. The classification names of difficult operatione areas should follow the categories in global surface cover data, such as Water Bodies, Wetland, Artificial Surfaces, etc. The classification of difficult operation areas needs to consider exploration and construction requirements and relevant policy restrictions.

According to the area of difficult operation areas, this study divides them into four levels: "$>$1 km$^2$, 1–0.1 km$^2$, 0.1–0.05 km$^2$, and $<$0.05 km$^2$". Difficult operation areas with an area greater than 1 km$^2$ have a certain distribution scale and are stable throughout the year, such as reservoirs, lakes, rivers, towns, etc., which have the greatest impact on oilfield exploration, deployment and surface engineering construction; The difficult operation area with an area of 1–0.1 km$^2$ has a significant impact on oilfield drilling construction, pipeline laying, and ground infrastructure, and should also be graded for consideration; The difficult operation area with an area of 0.1–0.05 km$^2$ also has a certain impact on oilfield drilling construction, as well as the construction of various factories and stations, which needs to be considered in gradation; However, the distribution range of difficult operation area with area less than 0.05 km$^2$ is relatively small, and work objectives can generally be achieved through avoidance in exploration, deployment and actual construction. Therefore, in this graded statistics, difficult operation areas with an area less than 0.05 km$^2$ will no longer be subdivided, and if necessary, layers can be hidden or not displayed to make the map beautiful.

After the above classification, the map of difficult operation areas can be overlaid with geographic information such as railways, highways, airports, place names, oilfield data such as drilling and mining rights, as well as policy restricted areas such as nature protected areas and ecological redlines. Finally, a comprehensive evaluation map of the ground environment of oilfield exploration areas can be developed (Fig. 5). This map not only comprehensively and objectively displays the types and distribution of difficult operation areas in the exploration area, but also accurately identifies, extracts, and analyzes the distribution of a certain or several types of difficult operation areas within a certain area. It provides a important and scientific work foundation and data support for the comprehensive evaluation of the ground environment in the exploration area, and can effectively support the decision-making of oilfield exploration deployment and mining rights evaluation.

Traditional ground evaluation is mainly based on the richness of surface data. Therefore, due to the limitations of data acquisition channels, it is difficult for most oilfields to obtain comprehensive and accurate surface information, and thus it is difficult to carry out accurate vectorization evaluation and classification research on difficult operation areas. This study is the first application of GlobeLand30 in vectorization evaluation of oilfield surface environment in China. The resulting map can fully and accurately display difficult operation areas on the surface, with advantages such as complete information, accurate location, and realistic morphology. After comparison, the error between the boundary position and the actual position of the surface features depicted by this method is less than 50 m. Considering the public welfare of GlobeLand30 and the need for oilfields to display all surface elements, this study only annotates the types of surface elements and does not involve issues such as place names and uses required by national confidentiality.

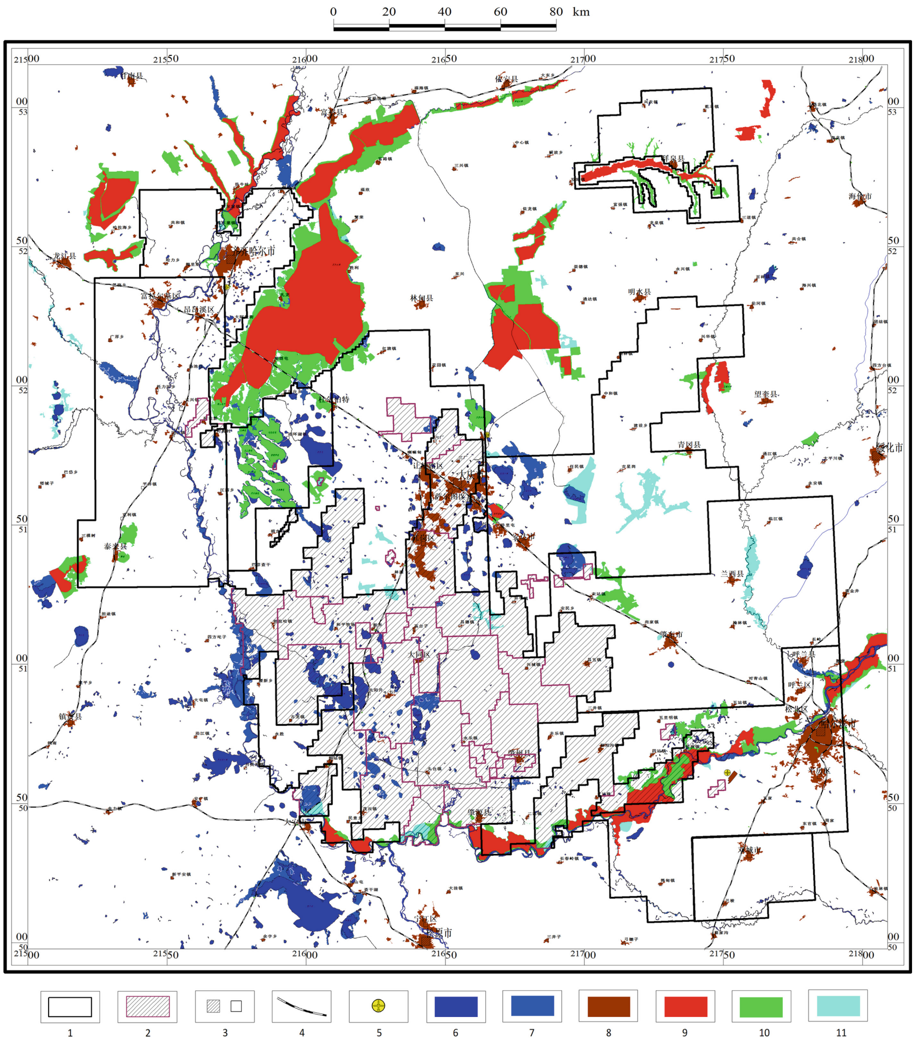


**Fig. 5.** Comprehensive Assessment Map of Difficult Operation Areas and Ground Environment in the Daqing Exploration Area in the Northern Songliao Basin. The types of difficult operation areas mainly include Water Bodies, Wetland, Artificial Surfaces, Nature Protected Areas and Ecological Redlines. The Artificial Surfaces mainly refer to the surface areas modified by human activities such as urban areas, railways, airports, etc. Nature Protected Areas and Ecological Redlines are policy restricted areas. Nature Protected Areas include core protected areas and general control areas. Legend Description: 1-Exploration Right, 2-Mining Rights, 3-Urban Area, 4-Railway, 5-Airport, 6-Water Bodies, 7-Wetland, 8-Artificial Surfaces, 9-Core Protected Nature Protected area, 10-General Control Nature Protected Area, Ecological Redlines.

## 5 Conclusion

This study introduces high-precision global surface cover data (GlobeLand30) into the exploration area of Daqing Oilfield for the first time, and establishes a ground environment assessment technology based on multiple information, providing comprehensive ground data information and technical support for the oilfield.

The vectorized ground environment assessment map has the characteristics of complete surface information and high accuracy, which can accurately determine the shape, location, and area of the "difficult operation areas". It not only greatly reduces the workload of ground exploration in the early stage of exploration deployment, but also provides a fine evaluation method for ground environment classification and grading, which has important guidance for the quantitative evaluation of the ground availability of exploration areas.

The ground environment assessment method based on GlobeLand30 has been promoted and applied in the exploration deployment and mining rights management evaluation of Daqing Oilfield, which has important significance and application value for guiding legal and compliant exploration, deployment decision-making, and medium-long-term planning scheme preparation.

## References

1. Department of Natural Resources, China. Opinions on several matters of promoting the reform of mineral resources management (trial), 31 December 2019
2. Ni, X., et al.: PetroChina reform and innovation practice of internal transfer and optimal allocation of mining rights and enlightenment. China Pet. Exp. **28**(1), 38–46 (2023)
3. Zhang, K., Miao, M.: Strengthening management of mineral rights blocks to promote reform in oil and gas industry. Sino-Glob. Energy **25**(10), 1–8 (2020)
4. Wang, X., Wang, G., Zhou, Z.: China national petroleum corporation promotes the internal assignment of the first batch of mining rights. China Petroleum Daily, 08 February 2017
5. Tang, G.-Q., Dong, X., Zhang, B.-S.: Assignment mechanism of mineral rights among giant state-owned oil and gas enterprises: review and proposals. Nat. Gas Ind. **39**(6), 147–155 (2019)
6. Li, G.-X., He, H.-Q., Liang, K., et al.: China's oil and gas resource management reform and innovative practice of PetroChina. China Pet. Exp. **26**(2), 45–54 (2021)
7. Chang, Y.-W., Liang, T., Zhao, Z.: The Trend of Oil and Gas. Petroleum Industry Press, Beijing (2017)
8. Liao, Y.-S.: The analysis on accuracy of geographical national census data and GlobeLand30 data. Surv. Mapp. **43**(5), 208–214 (2020)
9. Zhang, J.-H.: Comparative analysis of the mining rights and the red lines of ecological protection in Jianghan oilfield. J. Jianghan Pet. Univ. Staff Work **34**(1), 106–108 (2021)
10. National Catalogue Service for Geographic Information Homepage. https://www.webmap.cn/commres.do?method=globeIndex. Accessed 6 April 2023
11. Qin, Y., Hu, S., Zhang, X.: The conflict of rights and settlement mechanism of oil and gas mineral rights in China **428**(1), 69–72 (2012)

# Well Clustering and Reservoir Segmentation Based on Machine Learning Analysis to the Extracted Features from Multiple Well Logs

Yupeng Li(✉)

Beijing Research Center of EXPEC ARC, Aramco Asia, Saudi Aramco, Beijing, China
Yupeng.li@aramcoasia.com

**Abstract.** In conventional reservoir modeling, the well log curve shape information is usually lost when it is up-scaled. In this study, we implemented machine learning method to capture the well log curve shape information of each well and clustering those wells in target spatial domain. One state-of-art machine learning algorithm used in most of time series classification area is implemented for feature extraction. All the wells are grouped into different groups according to the similarity of extracted features from multiple log curves. The final spatial reservoir segmentation is obtained through spatial interpolation from the clustered well groups. The results of the 2D spatial map provides valuable insights to the depositional background analysis through providing the lateral geological heterogeneity features. One small synthetic data set is used in case study to illustrate this method as an effective way to characterize the spatial heterogeneity. As illustrated in the case study, the proposed spatial segmentation technique can be used as a fast geological modelling method to integrate geological heterogeneity features embedded in multiple well logs.

**Keywords:** Machine learning · Feature extraction · Multiple data integration · Geological heterogeneity characterization · Fast segmentation

## 1 Introduction

Recently, various machine learning (ML) algorithms have been successful implemented in petroleum exploration and development. It has been proved as an efficient and cost saving technique to harness the vast amount of reservoir-related information [1–5]. This article reports an innovative reservoir segmentation technique aiming to characterize the subsurface geological heterogeneity using high dimensional feature extraction, time series clustering, and spatial mapping.

In the proposed workflow, the well log curves of the target formation will be extracted as time series along the well trajectory. Particularly, Time Series Classification (TSC) methodology is used to group the target formation wells into different clusters or groups. In the TSC, it involves running a sliding window across each well log series. Then,

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 538–543, 2024.
https://doi.org/10.1007/978-981-97-0272-5_46

discretising the window to form some words, forming a histogram of word counts over the dictionary. Finally, it is to construct a specific group based on the obtained histograms.

Each group would represent a kind pattern in well log data domain. The study shows that TSC method can capture the stacking pattern of each well log curve. More importantly, the clustering can capture the vertical geological stacking patterns represented by each well log, which is essential for reservoir heterogeneity characterization or reservoir segmentation. In coming section, more details will be given to each step.

## 2 Methodology and Workflow

### 2.1 Time Series Transformation from Multiple Well Logs

In this paper, we seek to extract the well log data along trajectory as time series data. Conventional time series data is a sequence of data points indexed in time order. It tracks a sample over time and allows one to see what factors influence certain variables from period to period. Time series analysis are useful in way of finding how a given asset, security, or economic variable changes over time.

For well log data, given a target formation top and bottom, the measurements along borehole could also be looked as time series data. Comparing to traditional time based time serious data, here data are measured along depth with a even depth interval which is equivalent to time intervals in conventional time series analysis.

First, let's assume the wells in a reservoir is denoted as $\mathrm{W}_1, \mathrm{W}_2, \cdots, \mathrm{W}_\mathrm{N}$. For each well, it has multiple log curves, which are denoted as $\mathrm{L}_1, \mathrm{L}_2, \cdots, \mathrm{L}_\mathrm{M}$ as shown in Fig. 1. There are different well log types along the measured depth, such as Gamma Ray (GR), Spontaneous Potential (SP), or resistivity logs, etc. Then, each well log curve is observed as a multivariate time series given certain target formation top and bottom. Each well log time series id denoted as $d_{\ell 1}, d_{\ell 2}, \cdots d_{\ell \mathrm{D}}, \mathrm{D} = 1, 2, \cdots, M$. The whole well log data dimension for a reservoir formation under analysis is $\mathrm{N} \times \mathrm{M} \times \mathrm{D}$, as shown in Fig. 1. The objective is to cluster all the N wells based on M well log curves which will be looked as time series of D measurements along depth.

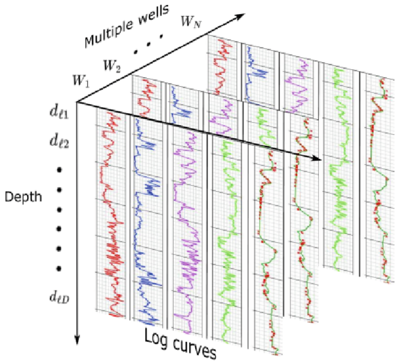


**Fig. 1.** Time series representation of well log data in a 3-dimensional form.

### 2.2 Time Series Classification with Dictionary Based Classifiers

In each time series, the curve has certain patterns. In classification, this shape pattern has to be captured. The dictionary approaches of Time Series Classification (TSC) address this by forming frequency counts of repeated patterns. In TSC, the dimensionality of time series are approximated and reduced by transforming those series into representative words. Then, comparing the distribution of words will calculate a similarity matrix between all of them.

More specifically, a window with certain size slides across each time series. Using this window sliding, it can form a pseudo-word using the information captured with the all the sliding window. In the algorithm, words are created using Symbolic Fourier Approximation (SFA), more technique detailed can be found in various reference [6, 7]. Then, it will construct a histogram of word counts over the dictionary. Usually, there is an encapsulates process that doesn't count trivially self similar words: if two consecutive windows produce the same word, the second one will be ignored actually. This greatly avoids a slow-changing pattern relative to the window size being over-represented in the resulting histogram. Finally, it will construct a classifier on the obtained histograms. Three approaches that have been published in the time series data mining literature are: Bag of Patterns (BOP); the Symbolic Aggregate Approximation Vector Space Model (SAXVSM); and the Bag of Symbolic Fourier Approximation Symbols (BOSS) [6, 7]. In this study, the BOSS method is implemented.

### 2.3 Clustering Ensemble and Reservoir Segmentation

After the wells are clustered in log data domain according to each log curve data, user will obtain M partition results from all the wells because each well log will have a different clustering result. Until now, the clustering or grouping of wells is done in latent space of well log curves data domain. The well locations are not play roles yet during clustering.

The spatial reservoir segmentation will count the spatial location into consideration. Reservoir segmentation is a procedure of spatial clustering using the ensemble clustering results based on the spatial locations. For the thickness of each target formation, it can be determined from the heterogeneity study requirement. Usually, for geo-modeling purpose, it could be just one zone that is believed to be the finest heterogeneity study vertical unit. Laterally, the principle follows the same rule of traditional geo-spatial statistics that the closer in spatial domain should have closer relationship in log curve latent curve domain. So, the traditional indicator facies modeling approach could be used. But, modeling indicator variogram is always a challenge [9, 10]. Thus, in this study, a conventional inverse distance interpolation method is implemented.

The proposed methodology workflow is illustrated in Fig. 2. For a whole reservoir, after detailed well correlation analysis, the segmentation will be done on each single zones. In each zone, the well log curves will be treated as a whole in clustering and segmentation using the workflow given above. Then, after segmentation is done for all the zones. The stacking of those segmentation results will be a full reservoir model for further heterogeneity characterization purpose, such as quick reservoir simulation or stratigraphy analysis.

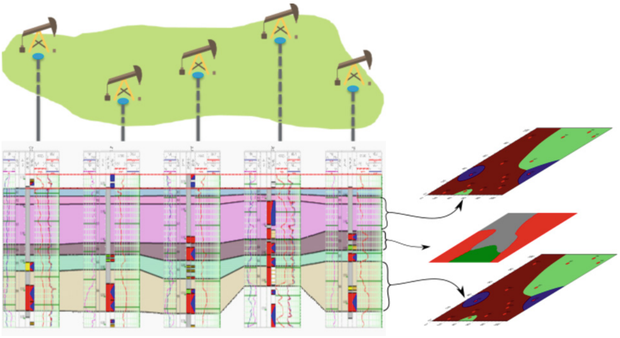


**Fig. 2.** Illustration of separately spatial segmentation given a certain target formation

## 3 Case Study

The proposed methodology is illustrated with one synthetic data set. In this data set, there are 13 wells in total as the well location map is shown in Fig. 3. For each well, only one zone is chosen as target analysis formation as shown in Fig. 3(a). The well location of the formation top will be used to do final spatial segmentation as shown in Fig. 3(b). Each well, there are nine well log curves available for clustering. For each well, the log attributes measured along the depth are transformed as time series for clustering as shown in Fig. 3(c). In practice, the entire wells trajectory are re-scaled so that the length of all the wells are the same given certain target formation top and bottom.

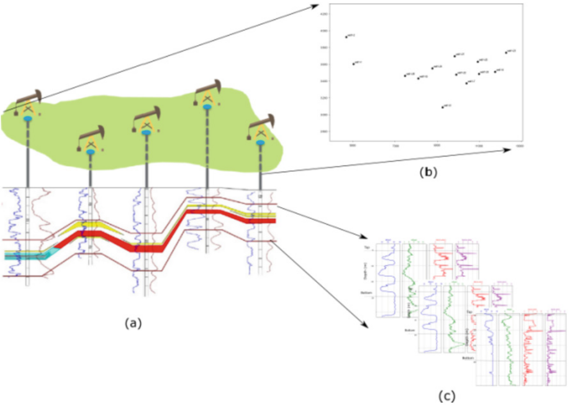


**Fig. 3.** The extracted time series from all the well log curves of the target formation

Using the proposed workflow, the spatial spatial segmentation is shown as Fig. 4. The final spatial segmentation result is consistent with those of different sedimentary environment for each well group based on the subsurface geological analysis. Thus, the obtained facies model can be looked as a facies or rock type modeling relatively to the facies modeling in conventional geological facies modeling.

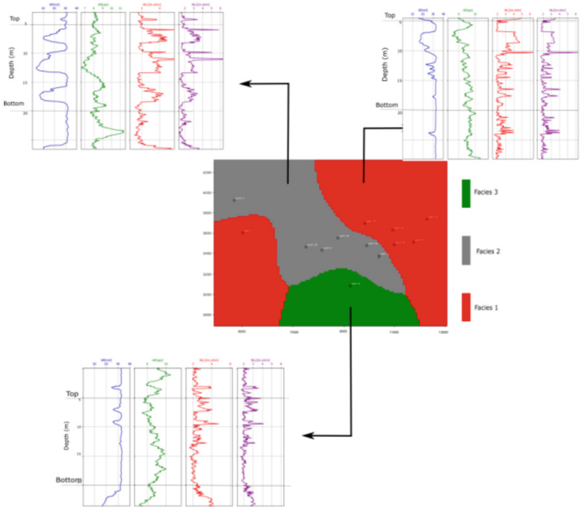


**Fig. 4.** Final clustering and spatial segmentation of the target area

## 4 Conclusion

It is an effective way to extract the geological feature through treating the well log data along well trajectory as time series data as illustrated in this study. The final obtained spatial reservoir segmentation can represent the heterogeneity lateral features revealed by the multiple well logs of the target formation.

The BOSS method can reveal the well log shape information from multiple log curves. Not only the well log measured values but also the stacking patterns, the shape message of each log curves, are accounted for in the time series analysis.

Alternatively, the 2D spatial segmentation obtained from each single layer or zone could be stacked together for a fully 3D reservoir facies or rock type modeling which actually can be used an effective fast geomodeling approach comparing to current existing modeling method.

## References

1. Asquith, G.B., Gibson, C.R.: Basic Well Log Analysis for Geologists. The American Association of Petroleum Geologists (AAPG), Tulsa (1982)
2. Archie, G.E.: Introduction to petrophysics of reservoir rocks1. AAPG Bull. **34**(5), 943–961 (1950). https://doi.org/10.1306/3d933f62-16b1-11d7-8645000102c1865d
3. Lis-Śledziona, A.: Petrophysical rock typing and permeability prediction in tight sandstone reservoir. Acta Geophys. **67**(6), 1895–1911 (2019). https://doi.org/10.1007/s11600-019-00348-5
4. Aliakbardoust, E., Rahimpour-Bonab, H.: Integration of rock typing methods for carbonate reservoir characterization. J. Geophys. Eng. **10**(5) (2013). https://doi.org/10.1088/1742-2132/10/5/055004

5. Sircar, A., Yadav, K., Rayavarapu, K., Bist, N., Oza, H.: Application of machine learning and artificial intelligence in oil and gas industry. Pet. Res. **6**(4), 379–391 (2021). https://doi.org/10.1016/j.ptlrs.2021.05.009
6. Schafer, P.: The BOSS is concerned with time series classification in the presence of noise. Data Min. Knowl. Disc. **29**(6), 1505–1530 (2015)
7. Bagnall, A., Lines, J., Bostrom, A., Large, J., Keogh, E.: The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Min. Knowl. Disc. **31**(3), 606–660 (2017)
8. Large, J., Bagnall, A., Malinowski, S., Tavenard, R.: On time series classification with dictionary-based classifiers. Intell. Data Anal. **23**(5) (2019)
9. Chiles, J.P., Delfiner, P.: Geostatistics: Modeling Spatial Uncertainty, p. 497. Wiley (2009)
10. Deutsch, C.V., Journel, A.G.: Geostatistical Software Library and User's Guide, New York, 119(147) (1992)

# Study on the Digitization Scheme of Gas Storage in Jidong Oilfield

Chong-Zhi Zhao[1], Cai-Yun Hu[1](✉), Yong Wang[1], Jia-Xi Fan[1], Wang da Lu[1], Yong-ke Hu[1], Zhi-feng Zheng[1], Xiong Liu[2], and Shi-Rao Wei[3]

[1] PetroChina Jidong Oilfield Company, Tangshan, China
np_hcy@petrochina.com.cn
[2] China Petroleum Kunlun Digital Intelligence Technology Co., Beijing, China
[3] China National Petroleum Construction Corporation Southwest Branch, Chengdu, China

**Abstract.** Continued low oil prices and the demand for green and low-carbon development have put forward urgent requirements for the traditional oil and gas industry to upgrade its production model. In recent years, with the deep integration of emerging technologies such as 5G, Internet of Things, artificial intelligence, cloud computing, big data, edge computing and the energy sector, new opportunities and challenges have been created for the traditional oil and gas industry to seek industrial upgrading, and digital transformation and intelligent development have become a way to achieve oil and gas fields Digital transformation and intelligent development have become the inevitable trend to achieve high quality development of oil and gas fields. As a guarantee unit for seasonal peaking and emergency gas supply in the Beijing-Tianjin-Hebei region, Jidong Gas Storage is responsible for the stable delivery of natural gas energy supply in the region. The non-linear relationship between gas supply capacity and formation pressure changes in natural gas storage reservoirs makes accurate regulation extremely difficult; the frequent changes of gas injection and extraction wellbore conditions make the control of wellbore integrity a IFEDC-202315047 2 great challenge; the non-linear changes of formation supply capacity in upstream storage reservoirs and the rapid changes of downstream customer demand make the ground system load fluctuate greatly, making it difficult to ensure the smooth operation of the pipeline network. In the face of these challenges, exploring the construction of an

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 544–565, 2024.
https://doi.org/10.1007/978-981-97-0272-5_47

intelligent gas storage reservoir with integrated and coordinated regulation of the ground, wellbore and formation is the best solution to meet this demand.

**Keywords:** Gas storage digitized · ground engineering · Digital delivery · 3D visualization · intergrated datamanagement

# 1 Present Situation of Intelligent Gas Storage Construction at Home and Abroad

## 1.1 Global Gas Storage Construction Status

As countries gradually shift from coal to clean energy natural gas in a large number of industrial and civil scenarios, and rapidly increase natural gas imports, how to safely reserve and timely dispatch natural gas is increasingly important. In summer, when the gas storage is traditionally filled, with the arrival of the winter heating season, the gas storage reserve will be emptied in winter, which involves the safe and stable operation mechanism of the gas storage.

There are 400 gas storages in the United States, 20 in Russia, Ukraine, Germany, Italy, Canada, France and China. There are many types of gas storage, of which the highest proportion of abandoned gas fields is 76%, the aquifer structure is 13%, the salt cavern is 6%, and the abandoned reservoir is 5%. The market-oriented operation mode of gas storage in Europe is relatively mature, all of which are operated by gas storage operators. There are 54 gas storage operators, including 14 operators in Germany, more than 4 operators in Austria, the Netherlands and the United Kingdom, and only one gas storage operator in 12 countries.

Since the outbreak of the war between Russia and Ukraine in February 2022, the EU has faced the risk of relying on Russian natural gas and Russia ’s reduction of natural gas supply, prompting the European heating season to fill the gas storage capacity to more than 80%. Germany is the EU ’s largest gas storage location, in this particular context by November reached 95% of the gas storage.

## 1.2 Construction Status of Foreign Intelligent Gas Storage

After decades of development, foreign gas storage construction technology is more mature. The intelligent function of gas storage focuses on the whole life cycle management and control of gas storage management. Starting from the numerical simulation analysis of target oil and gas reservoirs, the intelligent control decision is established by means of neural network to realize the intelligent production and injection allocation of gas storage, and gradually optimize the whole cycle operation and management mode of gas storage.

### 1.2.1 Czech RWE Company Intelligent Gas Storage Scheme

RWE Gas Storage CZ (RWE) is the largest natural gas storage operator in the Czech Republic. It has developed an integrated platform with Schlumberger [1]. The platform includes data center and gas storage engineering analysis function realized by data mining technology. It can automatically perform history matching to predict gas reservoir parameters, determine formation constraints, and predict maximum injection and production gas volumes. It can realize active intelligent gas storage management functions such as over-limit alarm, production prediction and optimization scheme push.

### 1.2.2 Dutch Bergermeer Intelligent Gas Storage Scheme

Research on dynamic mechanism model around digital twin is one of the important directions of intelligent gas storage. Based on the dynamic simulation technology of digital twin, the Bergermeer gas storage in the Netherlands has developed a decision support and prediction system [2], which provides a virtual model consistent with the field operation, predicts and intervenes in the production plan data, and predicts the future dynamic changes of the system in real time based on the process system to assist in supporting the field operation and decision-making.

### 1.2.3 Italian SNAM Company Intelligent Gas Storage Solution

SNAM, Italy 's national gas pipeline network company, independently developed the ARPOS system [3], which can display the entire gas storage system 360°, monitor operating parameters in real time and perform intelligent diagnosis and provide optimization solutions. The system integrates functions such as data visualization display, multi-level trend gas storage operation trend, and oil and gas well performance evaluation.

## 1.3 Domestic Intelligent Gas Storage Construction Status

The construction of gas storage in China started late. At present, the number of gas storage (group) has reached 17, 38 new gas storages will be added in the future, and the working volume will exceed 40 billion cubic meters. The construction of intelligent gas storage in most companies is also in the exploratory stage.

### 1.3.1 Hutubi Intelligent Gas Storascheme

Hutubi gas storage in Xinjiang is the largest natural gas storage in China.In the process of design, construction and operation of the gas storage, the Internet of things and big data technology are fully applied, and the functions of data acquisition, state monitoring, risk early warning and remote centralized control are integrated to realize the unattended operation of single well and distribution station, improve production efficiency and safety factor, and reduce operation cost. Through the construction of automatic production, digital office and intelligent management, the Hutubi gas storage has established a new management mode of single well unattended + regional centralized control + remote support and cooperation of control center. At present, one gathering station, three distribution stations and 45 single wells have realized the construction of digital stations. Build an intelligent office platform, accurately grasp the operation status of the production process through the platform, and make production adjustments through historical data analysis.

With the help of digital construction, Hutubi gas storage actively responds to extremely cold weather and increases supply guarantee. Technical personnel use intelligent office platform to carry out " one well, one policy " management of gas production wells, do a good job in fine analysis of geology, temperament and water quality, pay close attention to gas production wells and formation pressure changes, and ensure the maximum injection and production capacity.

### 1.3.2 Xiangguo Temple Intelligent Gas Storage Scheme

With the help of wired production network and IOT control technology, the Xiangguosi gas storage reservoir integrates numerical simulation models of gas reservoir, wellbore and ground on the basis of digital platform, uses advanced big data, cloud computing, artificial intelligence, image recognition and other technologies, builds an intelligent gas storage cloud platform supporting digital twin, establishes a three-in-one dynamic simulation system, divided into three levels: collection and injection station, gas storage management office and research institute The system is divided into three levels, i.e., gathering station, reservoir management office and research institute, and realizes the functions of production management, operation optimization, safety control, staff training and operation management in five aspects, i.e., “one platform”, “three levels” and “five aspects”. The aim is to realize the four intelligent functions of “comprehensive perception, automatic control, trend prediction, and optimal decision-making” [4, 5].

## 2 Implications of Digital China Construction Practice for Gas Storage Digital Transformation

In recent years, a new generation of digital technologies, represented by mobile Internet, cloud computing, big data, etc., has been deeply integrated with traditional industries such as transportation, finance, retail, manufacturing, etc., creating more high-quality new industries and new models, and playing an increasingly important role in the transformation of China’s economic growth from “high speed” to “high quality”. This has played an increasingly important role in the process of transforming China’s economic growth from “high speed” to “high quality”. At the same time, digital technology is gradually spreading from the economic field into other areas of practice such as government management and enterprise transformation, expanding horizontally and extensively. Innovations are emerging and fruitful, producing more successful cases in strengthening the people’s “sense of access” and enhancing the core competitiveness of traditional enterprises.

Industry, academia and research circles have carried out a series of exploration and research work around the basic logic of how digital technology can empower the change of traditional production mode, and have refined some more scientific guidelines to guide the implementation of digital transformation in the industry.

In his “Theory of Data Elements” [6], Rong Ke summarizes the newly emerged factors of production (land, labor, capital, knowledge, technology, management, data) in the evolution of the agricultural economy, industrial economy, and digital economy from the perspective of economic history, based on the assertion that “the contradictory movement of productive forces and production relations is the fundamental driving force for the continuous development of human society", In addition, it also compares data resources with traditional resources such as land, labor, capital, knowledge, and technology across history, elements, and forms, and then proposes specific ways to build a data ecology, discusses the importance of data elements for building a new development pattern and promoting high-quality economic development based on a new development stage, and also provides an opportunity for It also provides a reference way for enterprises to scientifically use data elements to upgrade their production and organization models.

In the period of global digital transformation, countries in Europe and the United States have introduced digital government innovation strategies, aiming to build an intelligent governance system led by digital governance. China also attaches great importance to the important role of government digital transformation in driving the modernization of national governance capacity. iShenzhen, Guangdong Provincial Affairs, Zheli Office, Yu Express Office, Digital Fujian and a series of digital technology-based government management practices, based on data integration and algorithms as the core, reshape the organizational form and responsibility system of government, public governance boundaries, governance capacity and governance technology, and government-enterprise interaction mechanism, etc., and have achieved remarkable results. The results are remarkable.

Central enterprises are generally larger in size and have more complex business processes, management decisions, organizational models, and industry chain synergies. More and more central enterprises are exploring the digital transformation of their enterprises and have released strategic plans or roadmaps for digital transformation. A preliminary statistical analysis of the digital transformation strategy plans of 62 central enterprises shows that at this stage, the focus of digital transformation of central enterprises in several fields is on strengthening data management. Central enterprises are committed to building complete, efficient and secure data management platforms and systems, implementing data sharing, mining data value, and providing scientific support for enterprise decision-making, thereby reconfiguring business models and empowering enterprise development. Specifically, the focus of the digital transformation of central enterprises is to take data management as the core, gradually realize the digital transformation of business and promote the upgrade of digital services and transformation.

In response to the practical achievements of digital transformation in various industries, the 19th Party Congress pointed out the important direction of building "Digital China", and in October 2019, the 4th Plenary Session of the 19th Party Central Committee established data as a factor of production for the first time; the 14th Five-Year Digital Economy Development Plan released by the State Council in 2021 further pointed out that data is the core engine of digital economy. In February 2023, the Central Committee of the Communist Party of China and the State Council issued the Overall Layout Plan for the Construction of Digital China, proposing the overall framework of "2522" for the construction of digital China [7–10], based on the application practices of various industries. The plan emphasizes that data is the core driver of enterprise transformation, and traditional enterprises should learn from the digital China construction plan, focus on data infrastructure construction and data resource integration, sharing and utilization, realize business data digitization, process digitization and decision digitization, and promote digital transformation; the overall framework of digital transformation is clarified, and the overall framework provided by the overall layout plan of digital China construction provides guidance and reference for enterprises to Digital transformation provides guidance and reference, emphasizing that digital transformation should focus on the adjustment of organizational structure and management system. The digital transformation of enterprises needs to focus on the adjustment of organizational structure and management system, and the concepts of digital governance, digital organization and digital talent proposed in the overall layout plan of digital China construction can provide reference for the digital transformation of enterprises.

# 3 Analysis of Intelligent Gas Storage Construction Strategy

## 3.1 Analysis of Gas Storage Business Management Activities

The purpose of intelligent gas storage construction is to optimize the business management capability of gas storage, improve the operational efficiency of gas storage, and reduce the operating cost and risk. Around this goal, figuring out the key pain points of the management of each business link of gas storage and introducing data technology to empower the business management process reengineering of gas storage is the key to do a good job of intelligent gas storage construction.

Compared with oil and gas field development and construction, the construction of gas storage reservoirs started late and the supporting technologies are still immature, facing many challenges, one of which is the complex geological conditions, broken tectonics, burial depths generally greater than 2500 m, and strong non-homogeneous reservoirs. The second is that the ground injection capacity is 5–10 times that of oil and gas wells of the same scale, and the block pressure of the reservoir is high and the components of the extractables are complex, so it is difficult to select and smoothly control the core equipment of the ground high-pressure large-scale injection and extraction during the operation period; the third is that it is difficult to analyze and complete the optimization of the reservoir injection and extraction operation from the aspects of market demand, block pressure difference, energy efficiency, etc.; the fourth is that the risk of safe operation is large. Fourth, the risk of safety operation is great. The alternate injection and extraction of large volumes of gas, pressure cycle changes are likely to cause geological instability, well control failure and ground equipment failure, how to real-time warning of formation, wellbore, ground facilities abnormalities, and prevent accidents is difficult. The traditional construction and production management model is no longer able to meet these challenges. The introduction of digital technology and the construction of intelligent gas storage suitable for the construction and production management model of gas storage in Jidong Oilfield will effectively promote the rapid, orderly, efficient and safe construction of gas storage [11–16].

## 3.2 Analysis of the Key Points of Intelligent Gas Storage Construction Management in Jidong

The accurate control of core business activities and the accurate control of core business management bottlenecks is fundamental to the good management of gas storage, and is also the main point of concern for the construction of intelligent gas storage. By summarizing the management problems encountered during the whole process of gas storage construction in the past two years, we analyze that the construction and operation of Jidong gas storage urgently need to improve the management efficiency in the following aspects:

### 3.2.1 Key Points of Management During the Construction Period of Jidong Gas Storage

During the construction period of the gas storage reservoir, it is necessary to optimize the sharing of the work results of the geological, engineering and ground design teams as well as the communication channels among the various departments involved in the construction of the gas storage reservoir, to strengthen the selection of equipment, and to enhance the efficiency of long-cycle equipment procurement to ensure the speed of gas storage construction.

Effective coordination among all parties involved in the construction of gas storage is the basis for the efficient construction of gas storage. The construction of gas storage is a comprehensive system project involving multiple professions and disciplines, which requires collaborative evaluation and selection in terms of reservoir capacity demonstration, compressor energy efficiency, comprehensive economic benefits, reliability design of drilling and completion engineering solutions, and design of ground engineering construction solutions. The research results of each profession support each other and there are mutual constraints, how to let each department in the process of professional research work in a timely manner to understand the dynamics of other professional work results and research index adjustment, timely feedback to the work of the professional statement to the brother professional response, will be able to efficiently mobilize resources to ensure the quality of gas storage construction and construction efficiency of the key. The research and construction period cross-professional research results sharing and standardized query platform, the construction of "engineering construction community of fate" digital ecosystem will be the most effective means to supplement the traditional management means across time and space management inefficiency.

Equipment procurement and selection is another key management point in the construction period of gas storage, and the cost performance, reliability and procurement efficiency of equipment selection are the main requirements in this stage. Due to the late start and small scale of the domestic gas storage construction, there is less research and development of equipment for the complex components of the gas storage and the production characteristics of "alternate injection and extraction, large injection and extraction, high pressure operation", and there is a lack of high-pressure large injection and extraction core equipment with superior performance in the market. Equipment selection, how to quickly grasp the dynamics of the market input of high-quality equipment in line with the operating conditions of gas storage, is the key to enhance the quality of gas storage construction; part of the construction of gas storage long-cycle equipment (6–12 months) whether timely delivery into the key factors that restrict the construction of gas storage can be put into operation on schedule. Under the premise of clear design requirements, how to realize efficient collaboration among construction units, operation units, material procurement units and bidding units, compress the technical capability assessment, price assessment, quality assessment, bidding and procurement cycle of long-cycle materials, save procurement cost and ensure equipment quality is the key management pain point during the construction period; the massive static parameters of equipment and construction process information are the important guarantee for equipment management and safety management during the operation period of gas storage. It is important to monitor and evaluate the leakage, corrosion, failure and process hazards of each key node of gas storage wells and ground, which need to be collected and collated from the design stage of gas storage scheme, and it is another key point of management during the construction period to realize the cooperation among all parties involved in the construction of gas storage and complete the data collection with high efficiency and quality. Therefore, the construction of a set of comprehensive control platform for equipment and facilities to solve the problems of supplier information assessment, cost performance evaluation of supply products, control of basic equipment parameters, detection and early warning of equipment operation indexes will effectively improve the quality of equipment procurement and selection management.

### 3.2.2 Key Points of Jidong Gas Storage Operation Management

The management objectives of the Jidong gas storage reservoir during the operation period are mainly focused on providing stable and reliable peak regulation and supply assurance capability, realizing a highly reliable integrity management capability of the trinity of geology, wellbore and ground during the operation of the gas storage reservoir, and exploring a set of highly efficient and high-level operation management mode.

Therefore, the focus of work during the operation period of gas storage has three main points. The first is to focus on the research of the digital twin model of gas storage reservoirs. Based on reservoir seepage research and wellbore-surface constraints, we will promote the study of underground and surface storage numerical simulation, and dynamically evaluate the technical indicators, safety operation parameters, and optimized operation parameters of each node based on reliable numerical simulation models. Second, focus on the optimization of operation and scheduling technology research. In view of the dynamic changes in demand for peak regulation and supply protection, as well as the bottlenecks in the production capacity of each node in the underground, wellbore and surface process equipment and facilities of gas storage, we will strengthen the construction of the production command platform, build an integrated production and scheduling platform with source data collection and multi-disciplinary coordination and integration that is adapted to the operational characteristics of gas storage, achieve a high degree of sharing of production source data, real-time control and real-time decision-making, and use information technology to realize the operational indicators of gas storage Automatic analysis and statistics, intelligent production and injection of injection and extraction wells, dynamic warning of wellbore safety of injection and extraction wells, monitoring and warning of ground equipment and facilities, and dynamic control to ensure efficient operation and management of gas storage, as well as to enhance the scientificity of operation and scheduling decisions; third, focus on safe operation and management of gas storage. Large-scale alternate injection and extraction of gas storage, pressure cycle fluctuations are prone to destabilization of the reservoir geological structure, well integrity failure and ground equipment failure, resulting in accidents. Jidong Gas Storage insists on scientific means to manage gas storage and digital means to improve the safety management efficiency of gas storage. By integrating the management requirements of three sets of technical documents, namely “Digital Delivery Specification for Oil and Gas Field Ground Engineering”, “Integrated Management Measures for Equipment and Facilities of Jidong Oilfield Company”, “Pipeline and Station Integrity Management” and “Integrity of Wellbore and Geological Body”, a set of above-ground and below-ground management documents is resolved. The data form of reservoir stratigraphic, wellbore and ground integrity management is integrated to meet the safety management requirements of gas storage construction and operation throughout the life cycle, and data collection and analysis are continuously carried out from the initial planning and design, project construction to the later production and operation of the whole life cycle to form data asset management, and based on the conclusion of data analysis, the risk evaluation of gas storage is done, remediation measures are formulated, and the construction of integrity system is promoted to ensure the safe operation of gas storage. To ensure the safe operation of gas storage reservoirs. The work direction includes geological integrity management, injection and extraction well integrity management, and surface equipment and facility integrity management. Geological integrity management focuses on monitoring whether the reservoir is sandy, the possibility of well wall collapse data, the production pressure difference of injection and extraction wells, and using the relationship between inventory and formation pressure to comprehensively evaluate and verify whether there is leakage in the reservoir. Injection and production well integrity management is mainly concerned with the three annular air pressure monitoring, annular fluid analysis, annular fluid level monitoring, wellhead corrosion monitoring and other work, the establishment of injection and production wellbore monitoring baseline, tracking annular air abnormalities. So as to achieve the purpose of reducing and preventing accidents and guaranteeing the safe operation of gas storage. The integrity management of ground equipment and facilities is mainly to improve the management of basic static data resources for ground process facilities and pipelines, standardize the collection and storage of regular testing and inspection data, determine the risk factors of gas storage through systematic analysis of the working process of each node of the ground system of gas storage, combine the results of statistical analysis of the causes of gas storage failures and accidents at home and abroad, and, based on the results of risk evaluation, target the existence of Based on the results of risk evaluation, preventive risk reduction measures are formulated and implemented to ensure the safe production of gas storage.

## 4 Intelligent Gas Storage Construction Plan

Referring to the “2522” overall construction framework construction idea of the “Digital China Construction Overall Layout Plan”, Jidong intelligent gas storage construction should pay attention to the “two basic” construction of digital infrastructure and data resource system at the early stage of construction, and at the same time of data resource construction improvement In the early stage of the construction of intelligent gas storage, Jidong needs to pay attention to the construction of digital infrastructure and data resource system, and at the same time, while improving the construction of data resources, Jidong needs to combine the key pain points of gas storage business management, and focus on the deep integration of digital technology with the management of gas storage construction process, the integrity management of gas storage stratum, wellbore and ground, the efficient operation and dispatch of gas storage, and the optimization of production dynamic analysis capability of gas storage, so as to create safe, efficient and high-level gas storage operation and management capability.

## 4.1 Overall Architecture

The overall planning strategy for the construction of smart gas storage is divided into two major phases according to the order of implementation:

The first stage is the construction of digital infrastructure and data resource system. Through the analysis of the business requirements of gas storage construction and operation management, the digital infrastructure needed for the construction of smart gas storage includes two aspects: first, it is necessary to focus on the construction of cost-effective and highly reliable wireless network system infrastructure to ensure that gas storage sites put in as many wireless sensors as possible to collect various operational parameters of gas storage [17]; second, it is necessary to tackle the highly concurrent IoT data storage and call solution to solve the problem of inconsistent data standards of various PLC systems and DCS systems in gas storage reservoirs, and to realize real-time and stable uploading of IOT data on data servers of each field station to the SCADA system in the data center of gas storage reservoirs [18, 19].

The data resource system construction focuses on sorting out whether the various types of data required for the efficient operation of gas storage reservoir business have achieved database control, whether the data sources are unique and reliable, and whether the data from various data sources can be integrated, etc. Through demonstration and analysis, the data base of Jidong intelligent gas storage reservoir needs to do a good job of building three major data resources, one is to do a good job of interconnection with A1, A2, A4, and Jidong oil field regional data lake, which can efficiently call The second is to build a static data resource base based on information collection of equipment and facilities and ground engineering construction process, to realize the control of all equipment and facilities of gas storage throughout their life cycle data, and to provide data support for the digital delivery of gas storage, integrity management of field stations and pipeline integrity management during operation; the third is to build an IOT database to realize The third is to build an IOT database to realize real-time data collection of equipment and facility operation parameters, provide data support for abnormal equipment operation warning, and provide reference basis for reasonable regulation and control of production operation. At the same time, it realizes real-time monitoring and early warning of dynamic changes in the whole process of gas storage and extraction (formation, wellbore and ground), and provides data support for the construction of intelligent applications for gas storage and digital gas storage. Based on the integration of the three data resources, a trinity data base of “surface engineering - wellbore - underground (gas reservoir)” is formed to support the construction of intelligent gas storage [20, 21].

The second phase focuses on the deep integration of digital technology and existing business management (intelligent application phase). Focusing on the construction of three information platforms to enhance the operation and management capabilities of gas storage reservoirs, one is the establishment of an information platform for the construction of gas storage reservoir ground projects, through the establishment of a unified platform for collaborative collection, collaborative processing, and collaborative release of project information to achieve the same source release and tracking of project results, major changes, and project progress plans at all levels, the refinement of information collection during construction, the standardization of management processes, the strengthening of project construction processes The second is to build a digital management platform for intelligent scheduling of production operations and dynamic analysis of gas storage reservoirs to achieve efficient and high-level operation and scheduling management; the third is to build a system integrity management platform to explore a new mode of integrity management that integrates oil and gas reservoir, wellbore, and ground, and to ensure the operational safety of gas storage. (See Fig. 1 for the overall architecture of intelligent gas storage).

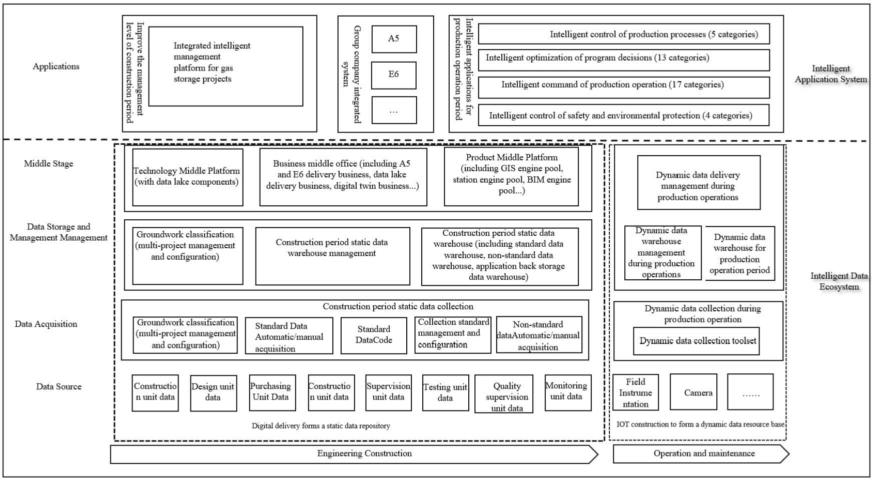

Applications
Improve the management level of construction period
Integrated intelligent management platform for gas storage projects
Group company integrated system
A5
E6
…
Intelligent applications for production operation period
Intelligent control of production processes (5 categories)
Intelligent optimization of program decisions (13 categories)
Intelligent command of production operation (17 categories)
Intelligent control of safety and environmental protection (4 categories)
Intelligent Application System
Middle Stage
Technology Middle Platform (with data lake components)
Business middle office (including A5 and E6 delivery business, data lake delivery business, digital twin business...)
Product Middle Platform (including GIS engine pool, station engine pool, BIM engine pool...)
Dynamic data delivery management during production operations
Data Storage and Management Management
Groundwork classification (multi-project management and configuration)
Construction period static data warehouse management
Construction period static data warehouse (including standard data warehouse, non-standard data warehouse, application back storage data warehouse)
Dynamic data warehouse management during production operations
Dynamic data warehouse for production operation period
Intelligent Data Ecosystem
Data Acquisition
Construction period static data collection
Groundwork classification (multi-project management and configuration)
Standard Data Automatic/manual acquisition
Standard DataCode
Collection standard management and configuration
Non-standard dataAutomatic/manual acquisition
Dynamic data collection during production operation
Dynamic data collection toolset
Data Source
Constructio n unit data
Design unit data
Purchasing Unit Data
Constructio n unit data
Supervision unit data
Testing unit data
Quality supervision unit data
Monitoring unit data
Field Instrume ntation
Camera
……
Digital delivery forms a static data repository
IOT construction to form a dynamic data resource base
Engineering Construction
Operation and maintenance

**Fig. 1.** General architecture of gas storage digitization

## 4.2 Data Ecosystem Construction Plan

The data ecosystem of gas storage is mainly divided into two categories, static data and dynamic data, static data from the unified A1, A4 and Jidong Oilfield data lake, digital delivery platform (new), dynamic data from the unified A2 and IOT data, through data linkage, import and other different ways to form the data ecosystem of gas storage (Fig. 2).

### 4.2.1 Static Data Resource Construction Program

Static data mainly refers to the design, procurement, construction and other engineering information data generated during the whole process of engineering construction, and its accurate control is necessary for good site integrity management and fine management

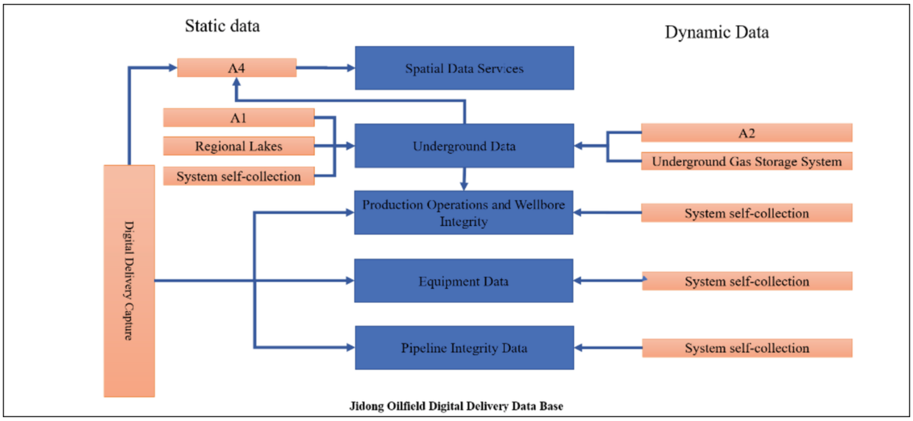


**Fig. 2.** Analysis of digital delivery data flow in Jidong Oilfield

of equipment and facilities, asset management and other business during the operation period of gas storage, the traditional management method is to store these data in the form of paper documents in the archives with the completion of the project, and the data guiding decision-making ability has not been given full play. Therefore, the best practical way to solve this pain point is to carry out digital delivery to realize the standardized management of engineering construction data.

Digital delivery mainly includes three aspects of work, which are to establish a ground engineering digital resource center that can be called efficiently, to establish a digital collaborative work platform, to establish a data visualization work platform, and to use digital delivery results and digital modeling technology to quickly establish a digital twin of engineering objects or entities in the physical reality world and to correlate with IOT data to further realize the operation of gas storage nodes in the The interaction and manipulation between the physical world and the digital world can realize automatic sensing, intelligent control, digital operation and intelligent management of gas storage operation (Fig. 3).

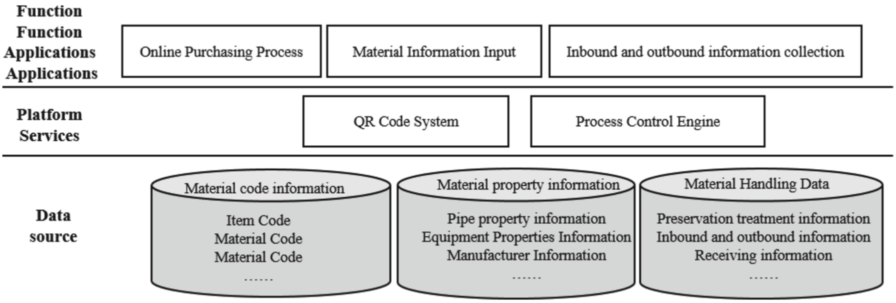

Function
Function
Applications
Applications
Online Purchasing Process
Material Information Input
Inbound and outbound information collection
Platform
Services
QR Code System
Process Control Engine
Data
source
Material code information
Item Code
Material Code
Material Code
......
Material property information
Pipe property information
Equipment Properties Information
Manufacturer Information
......
Material Handling Data
Preservation treatment information
Inbound and outbound information
Receiving information
......

**Fig. 3.** Material control scheme of Jidong Oilfield

#### *4.2.1.1 Construction of static digital resource center for ground engineering*

In the construction of static digital resource center for ground engineering of gas storage reservoirs, the key is to form and improve the digital delivery technology support system, mainly including: standardization of digital delivery, data collection process and template, support for standardization and standardization of source data collection; based on a unified platform, support for collaborative review and online evaluation of engineering design with the participation of multiple parties, and realization of online annotation and annotation of models and drawings It adopts mobile app technology to realize construction site data collection, automatic data acquisition and offline storage without network; through the implementation of the "Digital Delivery Technology Regulations for Oil and Gas Field Engineering of Group Companies" in Jidong Oilfield, it fully realizes standardization, standardization and online audit and archiving of engineering data. It supports the process approval of documents in the whole process of design, procurement and construction, and the online generation and archiving of archival documents.

By integrating the business management requirements of CNPC's "Technical Provisions on Digital Delivery of Oil and Gas Field Ground Engineering", "Pipeline and Station Integrity Management" and "Integrated Management of Equipment and Facilities of CNPC Jidong Oilfield Company", the data of ground engineering construction is standardized according to three dimensions of data type, data generating unit and data attribution business scope, and 527 data forms are resolved. Among them, the data are divided into structured data, semi-structured data and unstructured data according to data type; the data are divided into owner unit, design unit, procurement unit, construction unit, inspection unit, supervision unit, monitoring unit, quality supervision unit and digital delivery contractor unit according to data generation unit; the data are divided into pipeline and site equipment data according to data attribution business scope, and the pipeline equipment data are divided into valve, high consequence area Identification, pipeline appurtenances, pipeline components, welding seams, monitoring and evaluation, hydraulic protection, surrounding environment data, site equipment data is divided into dynamic equipment, static equipment, skid-mounted equipment, thermal equipment, fire equipment, self-control equipment.

#### *4.2.1.2 Material whole process control function module construction program*

The core of the construction of this function is to realize the effective collaboration between multiple participating units (design unit, construction unit, bidding unit, material supervisor, construction unit and suppliers) in the whole process of material procurement through QR code, so as to improve procurement efficiency and improve the quality of procured materials. The specific approach is to build online procurement process flow, through the process control engine, the material demand plan development, material procurement, delivery, storage, material clearance, installation and other processes to complete the online circulation, online collection; construction of material information collaborative collection, through the role to complete the material information entry, such as for the same bit number of materials, the establishment of sub-authority entry interface, design units to complete the design parameters collection, suppliers This avoids the traditional problem that each department only establishes data ledgers related to the business management of the department, and realizes the integration of material information; establishes a full-cycle monitoring program for material status, and through the whole process of QR code, users can grasp the overall real-time status of project materials from all levels to achieve the purpose of material tracking management. Achieve the purpose of material tracking and management [22, 23].

*4.2.1.3 Digital Collaboration*

Based on the standardized collection and management of data resources, digital collaboration enables the owner and all parties involved in the construction to participate online in the review of technical proposals (design, construction, etc.) and models at all stages, improving the review efficiency and realizing the closed-loop processing of problems; realizing the visualization of production operation parameters and on-site video monitoring. For example, the document distribution collaborative process is shown in the following figure VII Digital Collaboration (Document Distribution Process). The effect achieved through the realization of digital collaboration: realize online collaborative review of 2D drawings and problem pinning; realize online collaborative review of 3D models and problem pinning; realize online distribution and transmission of 2D drawing documents.

*4.2.1.4 Data visualization*

On-ground visualization realizes the association between 3D model and progress plan, and realizes dynamic control of project progress plan through 3D progress visualization simulation; meanwhile, the platform realizes the visualization display of key business index information based on 3D model and business comprehensive statistical information, provides intuitive and unified information display window for project management personnel and project visitors, and provides an auxiliary means for project management personnel to control the overall construction appearance of the project. Auxiliary means.

### 4.2.2 Dynamic Data Resources Construction Program

The dynamic digital resources required for the construction of intelligent gas storage mainly include production parameters and dynamic monitoring parameters collected by A2 system and real-time data such as pressure, flow rate, temperature, liquid level and vibration collected by IOT equipment, which are the basis for tracking the dynamic changes in production, warning the overrunning of production facilities and assisting the decision of injection and extraction scheduling. Among them, A2 system is a dynamic database that has been in mature application in the gas storage reservoir. Every day, the basic data reflecting the production dynamics are entered into the system class by the grassroots personnel, which is the basis for the dynamic analysis of the gas storage reservoir carried out by various departments of the reservoir, but the data collected by A2 system is coarse in granularity and not comprehensive enough to meet the real-time warning of the operation status of the injection and extraction wells and production field stations, and the abnormal state of the production facilities Therefore, it is necessary to build a plant-level industrial data resource center to provide unified standard data services for gas storage production management and production optimization to meet the needs of efficient development and management of gas storage in the future.

Focusing on the requirements of intelligent applications on data quality and data granularity, the focus of the construction of the plant-level industrial big data resource center is to consider the performance of the front-end intelligent instrumentation equipment and DCS system, the communication protocol and network transmission quality between the DCS and other regional data centers and the plant-level industrial big data resource center, the data reading and writing capability and data quality of the plant-level industrial big data resource center.

Among them, the intelligent instrumentation equipment of the field station and the construction of the DCS system, fully investigated the minimum requirements of the oil and gas reservoir analysis department, production operation management and other departments on the scope and frequency of data collection and data quality, combined with the later business needs, it is clear that the intelligent instruments attached to the injection and extraction wells all adopt wireless transmission, the transmission protocol is rola, the frequency of equipment collection can be set as required, and the minimum frequency of equipment collection is not In addition to the minimum requirements for DCS controller performance, controller operating load rate, processor switching stability, real-time switch acquisition, real-time network transmission and other indicators, and focus on the DCS system read and write performance, the DCS system is required to meet the normal business needs. The DCS system is required to meet the normal collection operation, and at the same time, it is also required to meet the higher-level system batch data call operation without crashing, and all new DCS systems of gas storage must provide OPC data transmission protocol to meet the data center can collect data of intelligent equipment of each field station through the standard protocol.

The network transmission link construction includes two parts, one is the communication between the individual equipment and the DCS system of the site, and the other is the communication link between the DCS system and the plant-level industrial data resource center. Among them, the communication link with the plant-level industrial big data resource center uses the company's existing fiber optic communication, using a dual link with one backup; the wireless communication between the equipment and the field station DCS system, taking into account the communication distance, communication cost, easy control of strategy, good scalability, minimizing the scope of failure and recovery time, reducing the overall TCO and other factors, using the wireless controller AC (wireless Access Controller) + thin AP (wireless access point) wireless network architecture, through the convergence switch and POE access switch, to achieve wireless terminal access and communication. Under this architecture, the wireless controller provides unified configuration and management for APs.

The plant-level industrial big data resource center requires the data server with data collection capacity above 100,000 points, plant-level data latency within 20 s, data concurrent writing capacity above 100,000 points, data collection and storage support distributed, meet the simultaneous access rights of 100 users, and the third-party applications can retrieve the data resources of the real-time data center in a stable and efficient manner. And the data server can meet the OPC and other protocol docking.

### 4.2.3 Above-Ground and Underground Data Integration Management Scheme

The above-ground and underground data integration management subsystem takes EPDM2.0 model standard and oil and gas field surface engineering digital delivery technology regulations as the integration service standard, and customizes the integration data service for each professional data on the basis of this integration service standard to support each system application.

The system application part mainly includes underground data management, aboveground data management, aboveground data visualization, and underground data visualization. After logging into the integration portal through the unified authentication service, users can enter each application module according to the assigned authority (Fig. 4).

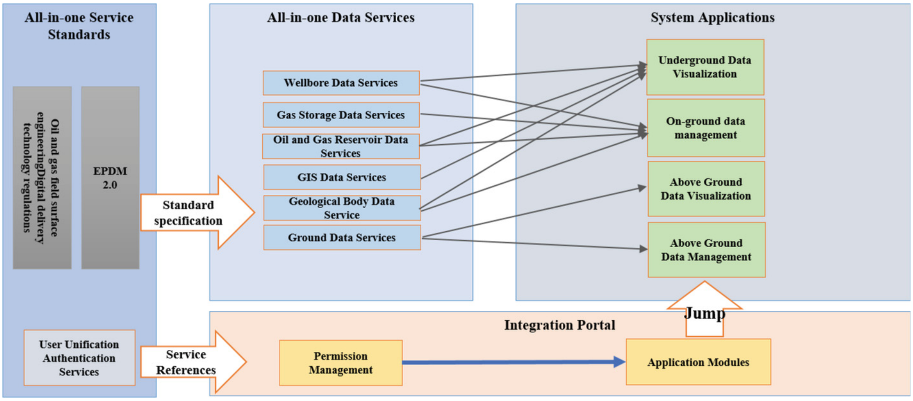

All-in-one Service Standards
Oil and gas field surface engineeringDigital delivery technology regulations
EPDM 2.0
Standard specification
User Unification Authentication Services
Service References
All-in-one Data Services
Wellbore Data Services
Gas Storage Data Services
Oil and Gas Reservoir Data Services
GIS Data Services
Geological Body Data Service
Ground Data Services
System Applications
Underground Data Visualization
On-ground data management
Above Ground Data Visualization
Above Ground Data Management
Jump
Integration Portal
Permission Management
Application Modules

**Fig. 4.** Integrated management of above-ground and underground data

### 4.3 Intelligent Application Ecosystem Construction Plan

Based on the reliable quality of data resources, and based on the business pain points exposed in the production management process of each section of the gas storage, the intelligent application ecosystem of the gas storage is formed by empowering the original production management process through information technology. The intelligent application ecosystem mainly focuses on the construction and management of gas storage ground engineering, the optimization of gas injection and extraction operation and scheduling management capabilities, and the construction of intelligent applications to improve the safety and production capacity of gas storage.

#### 4.3.1 Integrated Intelligent Management of Gas Storage Projects

The construction goal of the integrated intelligent management module for gas storage projects is to achieve the perfect project construction process data collection, online display of project promotion constraints, and unified release of project goals and project progress by building a shared information technology platform to eliminate the problem of untimely and inaccurate access to project construction information by all parties involved in project construction, which affects the overall construction efficiency of the project and improves Project construction management capability.

This function enables the creation, modification, application and sharing of projects in the form of user rights and application side to meet the needs of various positions in the enterprise and maximize the management efficiency. Taking a single project as a unit, it correlates information on data, drawings, progress, quality, safety, technology and cost in the construction building process, forming a digital management solution for engineering projects, providing data support for projects and realizing effective decision-making and fine management. The main functions designed are as follows:

First, project schedule management. By inputting project schedule plan and generating Gantt chart simultaneously, it visually reflects the deviation of project schedule and provides early warning for the project.

The second is the safety management module, including safety system, man-machine management, safety inspection, hazard sources, dangerous works, technical briefing, which are the key concerns of project safety. It insists on the policy of safety first, prevention first and comprehensive management, implements the work concept of safety management by everyone and safety grasp by everyone, cuts into safety management from various aspects, records the whole process of safety management, helps safety personnel to manage more easily and efficiently; realizes data trace, data sharing, accident warning, and prevention before it happens.

Third, the quality management module, including quality system, quality inspection, technical delivery. It helps project managers to ensure project quality in all aspects, from system establishment, to task responsible person implementation, to technical handover, to quality hidden danger investigation before it happens.

Fourth, the file management module. According to the project acceptance and archiving document specification required by oilfield ground engineering construction, standardized file management process. It contains function modules such as document reception, pre-assembled files, archiving management and basic settings, etc. It receives external electronic documents through local uploading of attachments, and realizes automatic assembling of files by setting file classification and assembling rules in advance to meet the whole process management from document reception, arrangement to final formation of electronic files.

Fifth is the intelligent site module. Using wireless network to integrate HD cameras, drones, mobile helmets and other video image data into the smart site platform to realize the functions of multi-departmental cross-regional remote supervision of construction site status, occasional spot-checking of construction situation, live and remote command of drones for major dangerous operations, and post-facto traceability of construction problems on the same webpage, and applying image recognition technology of deep learning algorithm to the construction site violation monitoring field [24] to solve the common construction violations with automatic warning and further promote the management quality of project construction sites.

### 4.3.2 Intelligent Production and Operation Management

The objective of the intelligent production operation management module is to establish a standard work page, integrate and display the work results among multiple departments, promote management synergy among multiple departments, realize digital enhancement of key operations such as geological research of gas storage reservoirs, injection and extraction regulation and control, transmission difference trend tracking, peak regulation optimization and wellbore integrity management, accurately control the changes of production indexes at each node, intelligently analyze production operation abnormalities and scientifically to make production control decisions. Based on this goal, the intelligent production operation management module includes a collaborative production operation and scheduling management module, a gas reservoir dynamics analysis module, and a gas injection and extraction formation and wellbore abnormality warning management module.

Among them, the goal of the collaborative production operation and scheduling management module is to realize the automatic production of technical reports required for operation management and dynamic warning of transmission difference changes. By calling real-time data from the Internet of Things and dynamic monitoring data in the A2 database, it realizes automatic generation, query, statistics and analysis of technical reports required for the operation and management of gas storage reservoirs to support the speed and quality of daily scheduling decisions for the production and operation of gas storage reservoirs; at the same time, by integrating real-time flow data and cumulative flow data from flow meters at each node, it visualizes the changes in the operation of gas injection and extraction of each gas storage reservoir as well as the changes in the transmission difference of each At the same time, by integrating the real-time flow data and cumulative flow data of each node, it can visualize the changes of gas injection and extraction operation of each reservoir as well as the trend of transmission difference of each reservoir, assist multiple departments to reach a unified understanding of transmission difference changes and take timely control measures in response to the trend changes.

The objective of the gas reservoir dynamic analysis module is to empower geological managers with information technology to improve the efficiency of injection and extraction well and reservoir dynamic analysis, and to improve the efficiency of dynamic assessment of reservoir capacity. The design functions include one-click generation and loading of basic data from common software for gas reservoir management and analysis, online generation of key technical charts for gas reservoir dynamic analysis, and early warning of abnormal parameters for gas reservoir operation [25].

The objective of the construction of the early warning management module for gas injection and extraction stratigraphic, wellbore, and surface abnormalities is to realize real-time dynamic monitoring of daily operating parameter abnormalities and to improve the safety management capability of gas reservoir operation. This is done by realizing real-time monitoring and early warning of formation and wellbore annular pressure data, and monitoring and early warning of surface sand emergence monitor data to realize the function of quantitative risk assessment of formation and wellbore integrity and to guarantee the production safety of gas wells.

### 4.3.3 Integrated Management of Gas Storage Equipment and Facilities

The goal of integrated management of gas storage equipment and facilities is to strengthen the ability of fine-tuned control of the whole life cycle of equipment and facilities, to enhance the ability of pipeline and equipment integrity management as the target orientation, and to build business function modules such as integrity data management, high consequence area management, and integrated management of the whole life cycle of equipment based on business requirements to achieve accurate control of gas storage equipment and facilities.

With the improvement of equipment refinement management requirements, the oilfield has an increasingly urgent need for static data consistency, traceability of management information, controllability of safety status, development and application of dynamic data and auxiliary decision support of various types of equipment. Through the construction of gas storage system integrity, it will achieve the standardization of gas storage equipment data and support the digitization of gas storage business and visualization of equipment management.

Equipment data standardization: Based on the static data transferred from the digital delivery platform, standardized data management of pipelines and equipment is realized; the digitally delivered equipment is classified and managed, and different classified equipment is provided with different attribute templates for management, so that equipment data is standardized and standardized. According to the concept of the whole life cycle of equipment, digital applications are established for each major link of equipment management, based on comprehensive evaluation of field stations, using IOT technology, with the goal of accurate operation and maintenance of equipment and dynamic control, planning to achieve automatic collection of equipment IOT information, remote control of key equipment, comprehensive coverage of security facilities, and standardized management level of equipment dynamic and static information to promote scientific and healthy management of equipment and ensure equipment In the process of gas storage construction and production operation to maximize the effectiveness; from decentralized management to unified management and from passive maintenance to active prevention.

Equipment-related master data integration: through the development of equipment basic ledger management function, build a unified equipment data collection portal and the ability to provide equipment ledger data service to other application systems; through the development of real-time production data management platform, form a clear “organization - production area - process equipment - real-time point The logical mapping relationship of “organization - production area - process equipment - real-time points” is formed; while maintaining the equipment ledger, the association between IOT data and the equipment itself is realized, which provides the basic conditions for the subsequent big data development and utilization scenarios such as production equipment fault monitoring and operation efficiency analysis.

Business digitization: through the high consequence area identification management module, realize the integrated management of the whole process of pipeline high consequence area identification information reporting, evaluation and release; realize the integrated management of the whole life cycle business of equipment and facilities, and reduce the workload of equipment and facility management.

Management visualization: realize pipeline visualization based on GIS map, associated query function, and static data information of pipeline related ontology and affiliated facilities by way of layers; combine with GIS map, realize hierarchical display of high consequence area location information and query display of related information; provide comprehensive statistical analysis charts of equipment multi-class data and data management kanban.

## 5 Application Scenarios

After the completion of the digitalization of gas storage, it will play a supporting role in the business management of the construction period and operation and maintenance period of gas storage business, and will be widely used in production management, operation management, geological drilling and extraction, ground engineering and other business scenarios (Fig. 5 Application scenario analysis).

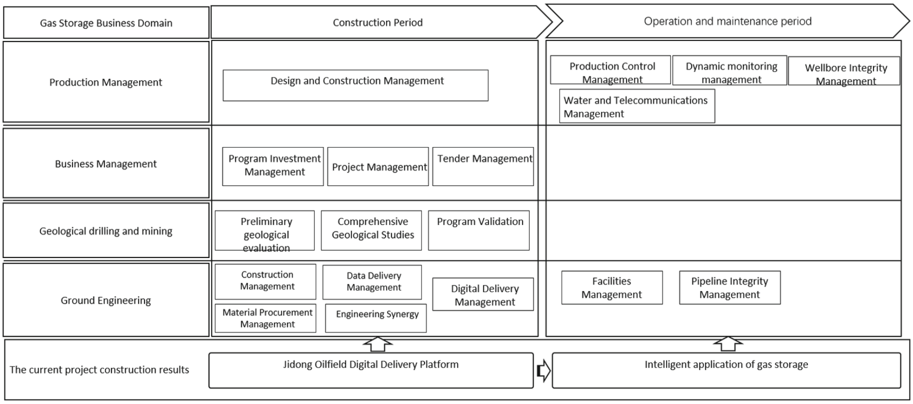

Gas Storage Business Domain
Construction Period
Operation and maintenance period
Production Management
Design and Construction Management
Production Control Management
Dynamic monitoring management
Wellbore Integrity Management
Water and Telecommunications Management
Business Management
Program Investment Management
Project Management
Tender Management
Geological drilling and mining
Preliminary geological evaluation
Comprehensive Geological Studies
Program Validation
Ground Engineering
Construction Management
Data Delivery Management
Digital Delivery Management
Material Procurement Management
Engineering Synergy
Facilities Management
Pipeline Integrity Management
The current project construction results
Jidong Oilfield Digital Delivery Platform
Intelligent application of gas storage

**Fig. 5.** Application scenario analysis

## 6 Research Conclusion and Outlook

The following conclusions are obtained through the research and analysis of the business pain points of the construction and operation management of Jidong gas storage reservoir.

Drawing on the experience of intelligent gas storage construction at home and abroad, the construction of an intelligent platform based on multi-data collection and fusion sharing, with digital twin as the core, to realize the collaborative control of gas storage stratum, wellbore and surface operations, real-time online monitoring and early warning of safety risks and technical indicators, and intelligent analysis and decision making will be a necessary path for digital operation of gas storage.

Intelligent gas storage construction needs to focus on planning to ensure the scientific nature of construction and cost performance. At present, the domestic construction of intelligent gas storage is in its infancy, the field is still a blue ocean, in many aspects of information technology construction work can achieve better results, but at this stage should be more cautious to promote the construction of information technology work, the need to systematically analyze the status of various aspects of the business management of gas storage, to catch the key pain points that restrict the level of business management of gas storage improvement, and based on this, to clarify what is the gas storage The construction of gas storage must be synchronized with the construction of information technology infrastructure, which is the key infrastructure can significantly improve the efficiency of the information technology construction of gas storage, which is the urgent need to improve the business management capacity of gas storage information technology construction functions.

Intelligent gas storage construction needs to do a good job of digital resource construction. Because the business management characteristics of gas storage reservoirs are different from the focus of traditional oil and gas field business management, the data system developed for traditional oil and gas field management cannot better meet the requirements of refined management of gas storage business, and in order to support multiple business disciplines to reduce the threshold and workload of information construction, it is recommended that in addition to improving the traditional oil and gas field data collection work such as A1 and A2, a new set of equipment and facility-based full In order to support multiple business disciplines to reduce the threshold and workload of information construction, it is recommended that in addition to further improving the traditional oil and gas field data collection work such as A1 and A2, a new set of static database for equipment and facilities with the goal of full life cycle management should be built, and a new set of real-time database with the goal of meeting real-time data call and big data analysis should be built to serve the efficient development of gas storage information construction.

Intelligent application ecology construction needs to focus on improving the ability of gas storage to regulate and maintain supply and safety production management. Only by meeting the core objectives of gas storage construction and the demand for safe production can intelligent gas storage construction be better realized. Unlike the traditional oil and gas field construction, the safe production and intelligent construction of gas storage puts more emphasis on enhancing the optimization of parameters operation in the working pressure range of gas storage; intelligent real-time completion of the adjustment of the production and injection distribution operation plan; the impact of the injection and extraction well switch of the production and injection distribution plan on the operation capacity and safety of the surface facilities; the integrity of wellbore, field station, pipeline and lean operation management. Therefore, the intelligent gas storage construction program should give priority to the construction of intelligent functions in the integrated scheduling of production and operation of gas storage for large injection and extraction, and the safety management of gas storage.

## References

1. Brown, K., Chandler, K.W., Hopper, J.M., et al.: Intelligent well technology in underground gas storage. Oilfield Rev. **20**, 4–17 (2008)
2. Xiping, Z., et al.: Top-level design of intelligent gas storage based on value chain analysis. Oil Gas Storage Transp. (04), 361–374 (2023)
3. Zhe, Z.: Enlightenment of foreign underground gas storage ground engineering construction. Petrol. Plan. Des. **28**(2), 1–37 (2017)

4. Haoran, L., et al.: Exploration and practice of intelligent construction of Xiangguo temple gas storage. In: Proceedings of the 32nd National Natural Gas Academic Annual Conference. Natural Gas Professional Committee of China Petroleum Institute: Natural Gas Professional Committee of China Petroleum Institute, China Petroleum Society Natural Gas Professional Committee, pp. 3005–3014 (2020)
5. Kanglin, Y.: The current situation and discussion of information construction of underground natural gas storage. Inf. Syst. Eng. (07), 124–126 (2019)
6. Rongke: Data Element Theory (2022)
7. Chunhui, W.: An in-depth interpretation of ' overall layout planning for digital China. Commun. World (05), 4–5 (2023)
8. The CPC Central Committee and the State Council: Overall layout plan for the construction of digital China. Future Urban Des. Oper. (03), 4 (2023)
9. Jiayi, Z.: The national cyberspace administration issued the ' digital China development report 2021.Technology China (09), 103 (2022)
10. The National Cyberspace Administration: Digital China development report 2022. People 's Post and Telecommunications (2023)
11. Lin, T., Gang, H., Bin, L., et al.: The cohesive force continues to innovate to create an efficient operation mode of gas storage. Petrol. Hum. Resour. (2), 43–50 (2019)
12. Ye, L., Xili, B., Nianrong, W., et al.: Ground engineering technology status and optimization suggestions of gas storage in China. Oil Gas New Energy **33**(5), 19–26 (2021)
13. Gangxiong, Z., Bin, L., Dewen, Z., et al.: Challenges and countermeasures of underground gas storage business in China. Nat. Gas Ind. **37**(1), 153–159 (2017)
14. Guanghua, Z.: Construction status and development suggestions of sinopec underground gas storage. Nat. Gas Ind. **38**(8), 112–118 (2018)
15. Ye, L., Gang, H., Lina, Y.: The challenges and countermeasures of gas storage construction in China during the 14th five-year plan period. Petrol. Plan. Set. **31**(6), 9–13 (2020)
16. Li Guoyong, X., Bo, W.R., Han, F., Yingbiao, W.: Suggestions on the layout of underground natural gas storage in China. China Mining. **11**, 7–12 (2021)
17. Hanrong, G., Dongqin, F.: The status and development trend of industrial wireless network. Chin. Instrum. (S1), 87–89+95 (2008)
18. Liang, M., Wei, W., Junli, W., Li, F.G., Ye, S.: Oil and gas field industrial big data construction and research. Inf. Syst. Eng. (10), 108–110 (2020)
19. Mingzhen, W., Rongquan, Z.: Research on the logic, challenges and countermeasures of industrial Internet platform driving high-quality development. In: Proceedings of the Second International Symposium on Frontiers and Practices of Public Governance Theory. College of Grammar, Department of Social Sciences, Yanshan University, pp. 229–235 (2022)
20. Guoliang, L.: Multi-platform big data integrated intelligent gas storage operation management system-taking H gas storage as an example. Sci. Technol. Innov. Bull. **08**, 70–71 (2020)
21. Kanglin, Y.: Present situation and discussion of information construction of underground natural gas storage. Inf. Syst. Eng. **07**, 124–126 (2019)
22. Xiaobing, T.: Design of safety management system for dangerous chemical gas cylinders based on RFID. Zhao Weidong. Fudan University, Tutor (2010)
23. Jie, Z.: Research on safety supervision system of hazardous gas cylinders based on electronic tags. Li Minbo. Fudan University, Tutor (2010)
24. Chongzhi, Z.: The operation identification method of valve specification based on Yolov5 s. Comput. Technol. Dev. **05**, 216–220 (2022)
25. Yuwen, Z.: Research on the innovation of 'four modernizations' construction and production operation management of petroleum enterprises. J. Econ. Res. **01**, 8–10 (2021)

# Construction and Application of Reservoir Dynamic Monitoring Data Management System Based on Cloud Primitive Architecture

Hong-mei Deng[1,2](✉), Liang Li[1,2], Wei-hua Yao[1,2], Yang Jiao[1,2], and Feng Li[3]

[1] Research Institute of Exploration and Development, Changqing Oilfifield Company, Xi'an, Shaanxi, China
denghm_cq@petrochina.com.cn

[2] National Engineering Laboratory for Exploration and Development of Low Permeability Oil and Gas Fields, Xi'an, China

[3] The Third Oil Production Plant of Petrochina Changqing Oilfield Company, Yan'an, China

**Abstract.** The original reservoir dynamic monitoring data management system of Changqing Oilfield was built in 2008. With the continuous development of oilfield, dynamic monitoring technology is also developing, business management is becoming more rigorous and standardized, and the operation problems of the original database are becoming increasingly prominent. Unable to meet and adapt to the actual needs of enterprise development, it also faces the bottleneck of iterative upgrade due to the difficulty of function expansion, which greatly hinders the subsequent digital transformation. System problems are mainly reflected in the following aspects: outdated system architecture, incomplete application functions, and inconsistent data standards and specifications. Therefore, it is urgent to reconstruct the original dynamic monitoring database system based on cloud native architecture and business application. Based on the development planning of digital transformation and upgrading of petroleum enterprises, this paper comprehensively expounds the construction and application of reservoir dynamic monitoring data management system based on H5 technology and oriented to micro service. Through the cloud upgrade of reservoir dynamic monitoring database, to solve the problem of timely, complete and accurate business data; Form closed-loop management of business chain including plan delivery, plan distribution, data archiving and plan tracking; By using the functions of report statistics, data analysis, graph drawing and comparison, we can better assist the management and technical personnel in application analysis, and guide the efficient development of reservoir by giving full play to the value of data assets.

**Keywords:** Dynamic monitoring · Cloud native architecture · Database · Data standard · Closed-loop management

## 1 Introduction

With the development of cloud computing technology, cloud native has become an important trend of enterprise information construction. As the core component of enterprise informatization construction, database needs to be optimized and upgraded in the cloud

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 566–571, 2024.
https://doi.org/10.1007/978-981-97-0272-5_48

native environment. The traditional database system has many problems in the cloud native environment, such as low resource utilization and poor Dr Performance. Therefore, the study of dynamic monitoring database system design based on cloud native has important theoretical and practical significance for improving the performance, reliability and scalability of the database system.

The dynamic monitoring system originally constructed in the oilfield adopts the C/S mode and is developed in the form of intelligent client. Due to the early development time of the system and the lack of structural upgrade in the operation process, the technical defects of the client are gradually exposed. In order to solve the security risks caused by the defects of technical architecture and further adapt to the requirements of current information development, this paper describes the comprehensive upgrade and optimization of dynamic monitoring database under the condition of cloud native technology. This paper mainly describes the construction and application of reservoir dynamic monitoring data management system of cloud native architecture.

## 2 Cloud Native Technology

Pivotal's Matt Stine pioneered Cloud native in 2013 with the Pivotal Cloud Foundry and Spring series of development frameworks. Cloud native architecture is an innovative software development method designed to make full use of the cloud computing model. Cloud computing platform uses cloud computing technologies such as virtualization to abstract physical devices into logical resources, connect multiple servers to form a larger resource pool, and improve resource utilization [1]. Cloud computing platform can be deployed in public cloud, private cloud and hybrid cloud, which is the basis of cloud-native technology.

The problem to be solved by the cloud native architecture is not simply to migrate the application to the cloud, but to maximize the non-business code part of the application through a set of architectural principles and design patterns, so that the cloud facility can take over a large number of non-functional features (such as elasticity, toughness, security, etc.) existing in the application, so that the business is no longer troubled by non-functional business interruption. It has the characteristics of light weight, agility and high automation (Table 1).

The concept of cloud native is to decouple software development using representative cloud native technologies such as container, Micro services, Devops and Continuous Delivery. On the one hand, it can improve the flexibility, easy maintenance and scalability of business development deployment. On the other hand, unified dynamic management and scheduling of applications and micro-services are carried out to improve work efficiency and resource utilization. The features mentioned above are not available in traditional monolithic applications, so the micro-service architecture is necessary to implement cloud native applications [2]. The cloud native composition (see Fig. 1).

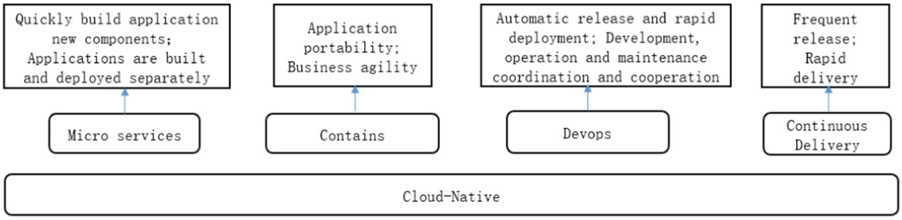

Quickly build application new components; Applications are built and deployed separately
Application portability; Business agility
Automatic release and rapid deployment; Development, operation and maintenance coordination and cooperation
Frequent release; Rapid delivery
Micro services
Contains
Devops
Continuous Delivery
Cloud-Native

**Fig. 1.** The cloud native composition

### 2.1 Container Technology

Containers are an important part of the cloud native concept. Containers are similar to virtualization, but unlike virtualization, they are a lightweight virtualization technology that provides users with a portable and reusable way to package, distribute, and run applications. The basic idea of a container is to package software that needs to be executed into an executable package. For example, package a Java virtual machine, Tomcat server, and application into a container image. Users can use this container image in the infrastructure environment to start containers and run applications, and can also isolate applications running as containers from the infrastructure environment. Containers are highly portable and users can easily run the same container in a development, test, pre-release, or production environment [3, 4].

### 2.2 Microservice

Microservices mean that different units or functions of the system run different containers, and the number of containers for each service can be adjusted according to its own load. For example, a large system includes user login, data query, application analysis and other functions, but all parts of the system do not linearly increase at the same time, some parts may be busy, and some parts may have more capacity [5, 6].

### 2.3 Devops

Devops literally stands for Development&Operations, Development&Operations. Devops emphasizes how the lifecycle management of software can be accomplished through automated tool collaboration and communication between efficient organization teams, allowing rapid delivery of development and stable deployment, and making the overall software development process faster by streamlining the process between Dev and Ops teams.

**Table 1.** The difference between traditional architecture and cloud native architecture

| Cloud native architecture | Traditional architecture |
|---|---|
| Predictable, cloud-native applications fit into a framework designed to maximize resiliency through predictable behavior | Unpredictable, often with longer build times, mass releases, scaling only gradually, and more single points of failure |
| Operating system abstraction | Dependent operating system |
| Resource scheduling is elastic | There are many redundant resources and lack of expansion ability |
| DevOps makes it easier for teams to collaborate | Departmental walls isolate teams from each other |
| Agile development | Waterfall development |
| Microservices are independent, highly cohesive, and low coupling | Single service coupling is severe |
| Automated operation and maintenance capability | Manual operation and maintenance |
| Quick recovery | Slow recovery |

# 3 Construction and Application of Reservoir Dynamic Monitoring Database System

## 3.1 System Design

With the continuous deepening of application, the problems of inconsistent data standards and specifications, imperfect application functions, and outdated technical architecture become increasingly prominent. In order to give full play to the guiding role of dynamic monitoring data, it is urgent to upgrade and improve the dynamic monitoring data model through unified data acquisition specifications, and continuously improve the quality of dynamic monitoring data. Improve application functions to better support business applications and assist management decisions; Through system architecture reconstruction, the system is scalable and easy to maintain to quickly respond to service requirements. For the actual business of oil and gas field dynamic monitoring, a database management application system is designed, which integrates dynamic monitoring plan, data acquisition, processing, storage management, and release and application of dynamic monitoring interpretation results (see Fig. 2).

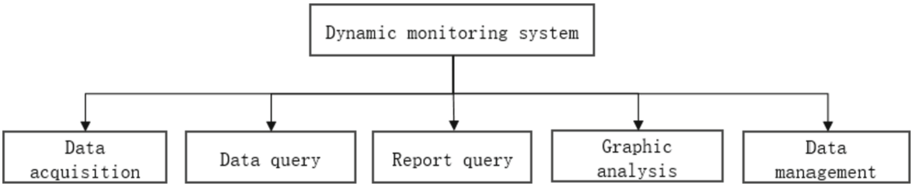


**Fig. 2.** Overall framework

### 3.2 System Construction

In this system, a standardized and complete oilfield dynamic monitoring database architecture has been established, and a data management platform and application release platform for oilfield dynamic monitoring have been built. The system construction is based on the original architecture of Dream cloud, and more than 50 micro-components have been formed. By means of “component-type development and block-type construction”, the cloud can be deployed quickly and iteratively. The whole-chain business process from test plan issuance, task distribution, implementation tracking and results submission is constructed to realize closed-loop management of dynamic monitoring business and meet the requirements of unified management and business application of oilfield companies [7, 8].

#### 3.2.1 Improve the Closed Loop Management Function Node

On the basis of upgrading the planned data delivery function of the original old oilfield, a series of functions from planned data delivery - single well work progress tracking - test data collection are developed for new stimulation and construction of single well. According to the current operation and management status of the production unit, the plan should be distributed in the factory level, divided into operation areas and blocks, and the functional links of the business process should be improved.

#### 3.2.2 Standardize the Dynamic Monitoring Data Management

The dynamic monitoring items, test technology names and archived data categories are standardized. The required data items and data filling rules of each type of monitoring items are sorted out and standardized. Establish data quality control rules, adopt two-level data verification mechanism, use data check tool for first-level verification, and data acquisition interface for second-level verification, to ensure the accuracy and integrity of data. The unified data collection portal supports multiple methods of batch import and single page input, realizing dynamic configuration of the collection module and enhancing system scalability. Manage the test team according to the bidding rules of three years to one field, and solve the problem that the name of the test unit is inconsistent, different oil production units correspond to the same test unit, and the data cannot be stored independently. To realize dynamic management of monitoring items and monitoring methods, the administrator can maintain the oilfield dynamic monitoring items and monitoring methods in a unified manner [9].

#### 3.2.3 Optimize and Improve Application Functions

In the report query, from the workload statistics, single well work progress, single well data step by step drilling query; Different dynamic monitoring methods can be distributed and displayed on the well map, monitoring projects can be queried by region, year and custom well, and legend attributes can be set flexibly. This paper focuses on the comparative analysis of original pictures and drawn pictures, which are urgent for business applications.

### 3.3 System Application

The system manages the dynamic monitoring data of 14 production units of the company, integrates the storage, management and analysis functions of dynamic monitoring data, and is widely used in oil field production management departments, scientific research departments and production departments. Through the interactive operation of the system platform, management and professional technical personnel at all levels can quickly and accurately monitor and diagnose the formation dynamic and production changes, so as to effectively guide the oilfield development and measure adjustment work, and play a positive role in the efficient management of oilfield dynamic monitoring business [10].

## 4 Conclusion

This paper mainly expounds the construction and application of cloud native technology and reservoir dynamic monitoring system, and deeply analyzes how to integrate information technology and business application. The reservoir dynamic monitoring data management system uses cloud platform (IaaS/PaaS) to build an information system and realize intensive and refined management of software and hardware resources. It adopts micro-service architecture to design the integrated application architecture of "platform + module", which can flexibly configure system functions, realize dynamic adjustment of posts, and support multi-post integration. The use of cloud platform and micro-service architecture can bring higher service quality, lower development, operation and maintenance costs, achieve rapid business iteration, and improve the efficiency of management decision-making and business analysis.

## References

1. Zhendong, X.: Design and Implementation of a Micro-Service Oriented Oil Data Dynamic Monitoring Platform. Xinjiang: Xinjiang University (2022)
2. Zhiming, Q., Liang, W.: Design of distributed database system based on cloud native. Comput. Eng. Design **40**(5), 1285–1290 (2019)
3. Xu Xiaohua, H., Zhongxu, C.F.: Architecture design and implementation of digital intelligence campus basic platform system based on cloud native. Mod. Comput. **21**, 117–120 (2022)
4. Chen, D.: Cloud computing database and travel smart platform design based on LSTM algorithm. Mobile Inf. Syst. **2022**, 1–9 (2022). https://doi.org/10.1155/2022/5124707
5. Yu Jingxin, D., Yong, S.W., et al.: Design and application of soil moisture monitoring system based on cloud native technology. Trans. Chin. Soc. Agric. Eng. **36**(13), 165–172 (2020)
6. Yun, W., Jianmin, Z., Junliang, L.: Design and research of database system based on cloud native. Comput. Eng. **45**(4), 1–5 (2019)
7. Jianjun, C., Xiaoqi, C.: Research on optimization of database system based on dynamic monitoring. Comput. Sci. **45**(3), 1–5 (2018)
8. Liren, Z., Ming, Z., Xiaoqi, C.: Research on performance optimization of database system based on dynamic monitoring. Comput. Sci. **44**(6), 1–4 (2017)
9. Jian, S., He, L.: Challenges and development of big data application in petroleum engineering. J. China Univ. Petrol. **36**(03), 1–6 (2020)
10. Xudong, C., Weidong, C., Xiaoyu, Z.: Application research of oil field production data management system based on B/S architecture. Comput. Meas. Contr. **26**(08), 142–146 (2018)

# Author Index

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 573–576, 2024.
https://doi.org/10.1007/978-981-97-0272-5