Prediction of Casing Damage: A Data-Driven, Machine Learning Approach

Yanhong Zhao
Department of Petroleum Engineering, China University of Petroleum-Beijing, Beijing, 102240, China

Hanqiao Jiang
Department of Petroleum Engineering, China University of Petroleum-Beijing, Beijing, 102240, China

Hongqi Li
Department of Petroleum Engineering and Data Mining Lab, China University of Petroleum-Beijing, Beijing, 102240, China

Received: November 6. Revised: December 14, 2020. Accepted: December 18, 2020.

***Abstract*—Casing damage is the result of a number of factors in the long process of oilfield development, so it must be correctly judged and repaired in time to ensure the normal production of the oil fields. With the development of data science, it has always been an imperative problem remained to be solved. In this paper, we adopt a data-driven and the machine learning approach to casing damage forecasts. Firstly, from the fields of geology, engineering and development, a lot of history data is collected and processed. Then, based on these dynamic and static data samples, the random forest algorithm is used to create the casing damage prediction model. Finally, after the model is tested in two fault blocks, the results indicate that accuracy rates are 91% and 75%, which proves the validity and performance of the mode.**

***Keywords* — Systems, System Science, System Engineering, Control, Casing Damage, Casing Damage Prediction, Data Driven, Data Governance, Machine Learning.**

I. INTRODUCTION

At present, as most oilfields enter the late stage of development, casing aging becomes severe and casing failures like damage and deformation frequently occur. Taking an oil production plant in Daqing as an example, as of December 2017, there were 1,147 casing damage wells. Among them, the casing damage in Pubei Oilfield is relatively severe, accounting for 54% of all casing damage wells, and the cumulative rate of casing wear is about 30%. The types of casing damage are mainly divided into deformation and rupture. About 64% of casing damage points are distributed in the Putaohua oil-bearing layer. Casing damage wells severely restricts the production of the oilfields, destroys the injection and production system of the oilfields, causing an unbalanced supply-discharge (injection and production) relationship and increasing the repair cost of casing damage. Meanwhile, casing damage also results in abnormalities of formation pressure and geological structure, which in turn induces new casing damage. Casing damage has become an unavoidable major problem for stable and high oilfield production. Casing damage is caused by a variety of factors during the development of the oil field, such as geological, engineering and development factors . Therefore, it is of great significance to research into how to determine the main control factors of casing damage in different blocks and develop a suitable method for the casing damage prediction to boost the development of waterflooding in blocks and make the adjustment to it.

The theories and research findings regarding casing damage at home and abroad have flourished and been widely discussed[1], such as "casing damage caused by mud shale flooding ", "casing damage caused by pore pressure difference ", "casing damage of oil reservoir caused by sandstone vertical deformation during high-pressure water injection ", "casing damage caused by in-situ stress concentration ", etc. The corresponding research methods include numerical simulation method , finite element method , etc. However, these researches on casing damage mechanism are limited and independent because they are mostly qualitative researches and study a single technology or a single factor. In addition, the factors that affect casing damage are non-linear, uncertain, and time-varying, making existing research methods and calculation models unable to accurately or efficiently predict the risks of single-well casing damage. With the development of big data and artificial intelligence technology, oil explorers and developers also keep exploiting big data technology to make the analysis and prediction of reasons for casing damage possible[2-8]. Yan Xiangyong et al. adopted a support vector machine approach to build a casing string life prediction model for oil and gas wells under complicated conditions with 32 casing failure factors as input vectors and the remaining life of the casing string as the output vector. Zhang Jie , Wang Liyan, etc. selected 10 indicators such as faults, shale content, porosity, corrosion perforation, and fracturing times as a factor set, and used a fuzzy comprehensive evaluation method to establish a casing damage prediction model, making it possible to quantitatively evaluate casing damage of oil and water wells. Based on the analysis of geological factors that affect casing damage, Jiang Xueyan et al. selectively chose four single indicators and introduced the deterministic coefficient to quantify and stack them in the same interval, obtaining the casing damage risk degrees of the geological factors that can be used to evaluate the risk of casing damage. Zhang Xu et al. adopted Bayesian neural network method which has higher accuracy compared with BP neural network and selected 14 parameters such as production time, wall thickness, steel grade and oil pressure as network inputs to predict the casing damage of oil and water wells. Huang Jun et al. comprehensively analyzed various factors and established a genetic neural network model based on the analysis of main components to predict casing damage of oil and water wells.

In summary, the machine learning methods based on artificial intelligence can provide a scientific basis for the prevention and control of casing damage wells to a certain extent due to its good capabilities of data fitting and non-linear modeling. By this method, various casing damage prediction models such as support vector machines and neural networks are established and the interaction among various factors can be taken into comprehensive consideration. Nonetheless, oil data covers multiple fields, such as geology, well logging, engineering, and oil development and spans a wide range of time and space. The cross-domain data fusion poses a serious challenge to the construction of machine learning datasets and also hinders the application of machine learning technology in the petroleum field. This paper proposes a set of data-driven casing damage prediction methods, including the process of constructing big data of casing damage, governing data, generating samples, and constructing and applicating prediction models as shown in the figure 1 below. First, lots of historical data is collected and processed from the fields of geology, engineering and oil development to build a big database for casing damage; then existing data is combined to establish casing damage dynamic and static samples in units of time and stratum, respectively. Based on these samples, identification models of casing damages in different horizons, risk identification models of casing damages in different blocks and prediction models of single-well casing damage are constructed by means of random forest training and fitting. Finally, two blocks are selected for operation to verify the prediction. The results prove that the method proposed in this paper has a good application and promotion value.

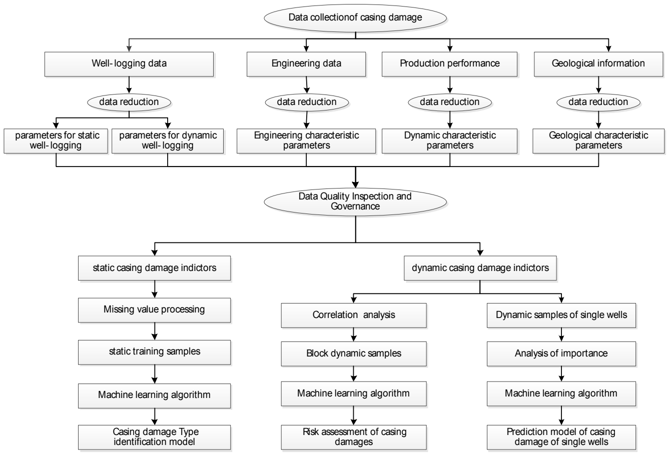
Data collectionof casing damage
Well- logging data
Engineering data
Production performance
Geological information
data reduction
data reduction
data reduction
data reduction
parameters for static well- logging
parameters for dynamic well- logging
Engineering characteristic parameters
Dynamic characteristic parameters
Geological characteristic parameters
Data Quality Inspection and Governance
static casing damage indictors
dynamic casing damage indictors
Missing value processing
Correlation analysis
Dynamic samples of single wells
static training samples
Block dynamic samples
Analysis of importance
Machine learning algorithm
Machine learning algorithm
Machine learning algorithm
Casing damage Type identification model
Risk assessment of casing damages
Prediction model of casing damage of single wells

Fig. 1. Data-driven prediction process of single-well casing damage

II. Data Construction for Casing Damage

A. Data Collection

Influencing factors of casing damage include engineering, geology, and development factors, cover geology, drilling, well-logging, well-cementing, well-completion, perforation, development, testing, fracturing, acidification and other fields and also involve more than 20 business data as shown in the figure 2 shown. According to data's time characteristics, the data is divided into basic data, static data and dynamic data. Basic data includes basic information related to blocks, layers, wells, and casing damage wells and so on. Static data is related to the stratigraphy, including data about well-logging, lithology, physical properties, perforation, and sedimentation, which can be used to analyze longitudinal distribution characteristics of casing damage. Dynamic data is related to time, including data on oil and water well production, acidification, fracturing, oil and water well pressure measurement, injection profile, and fluid production profile, etc.

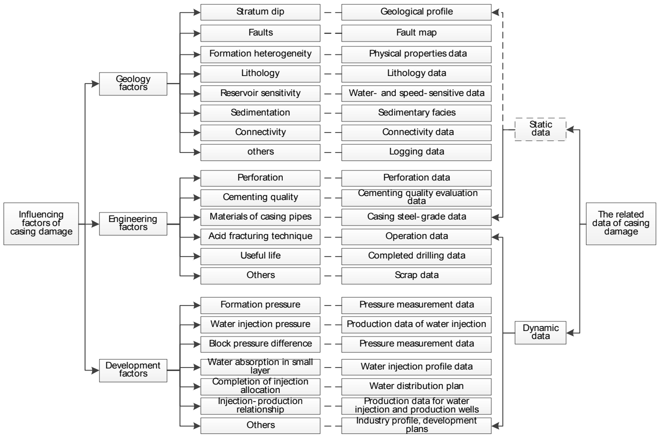
Influencing factors of casing damage
Geology factors
Engineering factors
Development factors
Stratum dip
Geological profile
Faults
Fault map
Formation heterogeneity
Physical properties data
Lithology
Lithology data
Reservoir sensitivity
Water- and speed- sensitive data
Sedimentation
Sedimentary facies
Connectivity
Connectivity data
others
Logging data
Perforation
Perforation data
Cementing quality
Cementing quality evaluation data
Materials of casing pipes
Casing steel- grade data
Acid fracturing technique
Operation data
Useful life
Completed drilling data
Others
Scrap data
Formation pressure
Pressure measurement data
Water injection pressure
Production data of water injection
Block pressure difference
Pressure measurement data
Water absorption in small layer
Water injection profile data
Completion of injection allocation
Water distribution plan
Injection- production relationship
Production data for water injection and production wells
Others
Industry profile, development plans
Static data
Dynamic data
The related data of casing damage

Fig. 2. Data related to influencing factors of casing damage

B. Data Quality Inspection and Governance

Casing damage data are distributed in different professional databases and its amount is huge. Moreover, database systems have been constructed in different periods, thus causing problems such as inconsistent information standards, diverse data types, different storage formats, duplicated data, and inconsistency between the old and new data. Since the data is an important resource for enterprises, its quality directly affects its value. Therefore, it is necessary to perform quality inspection and management activities on the collected raw data to improve the data quality and reduce the impact of "dirty" or "bad" data on data analysis and data mining. During the process of building big data for casing damage, data quality inspection includes integrity inspection, consistency inspection, accuracy inspection and validity inspection.

Integrity inspection. The degree of difference between the data that you want to collect and that you actually collect can be gained by means of data integrity inspection. When constructing samples of machine learning, you can choose wells with better data integrity for modeling. This paper uses two-dimensional matrix to represent the integrity of the casing damage dataset where a row represents a well, and a column represents the data category, such as layered data, small layer data, logging data, and so on. Cell [i] [j] = 0 means that the j-type data of the i well does not exist; data [i] [j] = 1 means that the j-type data of the i well exists. If the data exists, the cell is filled with green and it is filled with red if it is a casing damaged well.

Consistency inspection. For any oil field in the long-term exploration and development process, there are inevitably lots of inconsistencies during the data integration process due to different database standards, coding methods and naming methods applied in different periods and so on. They are mainly reflected in the inconsistencies of horizon information, time formats, names of well-logging curve, logging instruments and calibration, etc. as shown in Table 1.

Table. 1. Inconsistencies and governance methods for related data

Problems	Descriptions of problems	Governance methods
Horizon	Layers data: N2t, K1n4, etc. Breakpoint data: K1n1 top, K1n2, etc. Sand-layer data and perforation data: represented by the group name of oil layer and small layer number, such as P1, 110, 021, 070. Lithology and property data: represented by sample depth.	Establish standards of horizon descriptions and unify information on horizon descriptions on the basis of the relationship between large- and small- layers and depth. Time formats.
Date format	2010-01-12, 198001, 1999/01/23 and other formats	Unify time format.
Names of Well logging Curve	The naming of logging curves is not uniform, which increases the workload in the course of processing logging data. Moreover, the names of the measurement items are not standardized, for instance, the naming of caliper is various: CAL, CALI、CALM, CALS, CALS1, CALD, CALX, CALY, and so is the naming of measured depth: DEPTH, DEP,etc.	Establish standard names of logging curve and mapping rules, and standardize the original names of logging curves.
Well logging instrument and calibration	Different logging instruments, operating methods and calibrators with different standards in different periods lead to errors in the measurement data of each well	Standardized processing of measurement data about single and multiple wells. This paper adopts histogram normalization/standardization method.

Validation inspection. The main task of data validation inspection is to identify invalid values, which include vacant values, duplicate values, infinity, infinitesimality, and special values agreed by different information systems, such as -99999 in well-logging data. Data loss is inevitable in various database systems, and each type of loss will have different effects on statistical analysis. In order to analyze the mechanism of data loss and evaluate the impact of data missing, the quantity and distribution patterns of missing data need to be identified. As is one of the focuses of data scientists, the detection and governance of vacant values usually requires to figure out the proportion of missing data, the correlation between missing data, or its correlation with observable data and confirm whether the missing data is concentrated on a few variables as well as whether it is widespread and randomly generated. The methods for vacant values inspection include list detection method and the graphic survey method. There are generally two ways to deal with missing data. One is interpolation and the other is deletion. Different mechanisms of data loss will make approaches to governing data loss different. If the missing data is concentrated on several relatively less important variables, you can delete these variables; if a small fraction of the data is randomly distributed throughout the dataset, you can consider imputing the missing data by means of proximity interpolation, mean interpolation, probability interpolation, maximum likelihood estimation of normal distribution data, multiple interpolation, which are all commonly used.

Accuracy inspection. The key to the accuracy inspection is to find outliers in the data. Outlier detection methods are broadly divided into two categories. One is outlier detection, such as box plot test outliers, clustering, and local outlier factor method Etc. the other is the detection of data within a reasonable range, using professional background knowledge to detect abnormal points. Taking petrophysical measurement data of well logging as an example, the values of different blocks and horizons all have certain distribution ranges, so those values out of their own distribution range values are outliers. In contrast, the second method is more reasonable in that not all abnormal values found with the first method are necessarily outliers, while a value out of range must be an outlier. For example, a point with a less than or equal to zero resistivity must be an abnormal point. Generally speaking, vacant values are also outliers. Therefore, the methods for managing vacant values are also applicable to management of outliers. With different application scenarios, the methods of detecting and managing outliers are unlimited.

III. CASING DAMAGE IDENTIFICATION MODEL

The casing damage dataset is a non-linear, complex, and geophysical system that spans a wide range of space and time. Only raw data is not enough to build machine learning samples. How to design relevant indicators to characterize different influencing factors of casing damage is the key to the construction of oil big data samples.

A. Static Samples Construction

Based on the statistical results of the casing damage dataset in Pubei Oilfield, geological and engineering factors affecting casing damage are observed to design static indicators such as lithology, perforation, sand layer, sedimentation, connectivity, and adjacent wells as shown in Table2.

Table. 2. The static factors for Pubei OilField

Factor name	Symbol	Distribution characteristics
Stratum dip	dcqj	The inclination angle is relatively low, with a slightly steeper inclination angle of 5 degree on the west wing, a gentler inclination angle of 3 degree on the east wing, 2 degree at the northern end, and an inclination of less than 1 degree at the southern end extending further.
Occurrence rate of faults	dcsf	About 22% among all wells encountered faults and 24% of casing damage wells encountered faults.
Casing damage rate of adjacent wells	lj	Different fault blocks are distributed in different districts. Taking a certain fault block as an example, the average casing damage ratio is 50% among adjacent wells within 400 meters.
The types of sedimentary facies	cjx	The sedimentary facies where about 65% casing damage point is located are mainly non-sheet sand, main channel and main sheet sand.
Lithology	yxlx	Most casing damage points are Mudstone.
Sand-mudstone interface distance	yxdis	Most of the casing damages are located within 1 meter from the sand-mudstone interface.
Perforation layer	sksf	About 75% casing damage points are in the perforation layer has and about 25% in the non-perforation layer.
Outer diameter	wj	The outer diameter of 90% casings is 140mm, but about 31% of casings with an outer diameter of 114mm experience casing damages.
Wall thickness	bh	The wall thickness of casings is mainly 7.7mm, but about 36% of casings with a wall thickness of 6.4mm suffer from casing damages.
Casing damage frequency	tscs	80% of casing damage wells experience casing damage once, 16% twice, 3.3% three time, and less than 1% four times.
Service life	sysm	First casing damage is18 years on average;

Lacking of data values is one of the problems often encountered in data analysis. Without high-quality data, there won't be high-quality findings obtained from data. There are three main types of methods for processing missing values: delete tuples, complete data and leave data aside. If the proportion of missing values is small, it can be discarded directly. if it is relatively large, deletion is not advisable for a lot of information will be lost in this way, causing a systematic difference between the incomplete observation data and the complete observation data. Analysis of such data may lead to wrong conclusions. Data completion is usually based on statistical principles, and fills a missing value according to the distribution of values of other objects in the initial dataset, such as average value filling, special value filling, regression replacement, etc. However, the filling of empty values may not be completely in line with objective facts and incorrect filling of null values might make things worse and cause incorrect results to be produced from data. Therefore, to deal with missing values requires detailed analysis of issues with their own uniqueness taken into consideration. Missing values should be derived and filled by using professional methods combined with their practical application scenarios to reduce the gap between machine learning algorithms and practical applications. In this study, the main reason for a large number of missing static indicators of casing damage is the incompletion of well-logging curves, which provide the basic data for calculating the physical parameters of the reservoir, for example, the integrity degree of curve AC is only 68%. For the missing data of well-logging curves, regression methods such as support vector machines and neural networks can be employed to establish data models of the acoustic curves so as to obtain estimations of missing data of the logging curve. After processing the missing values, the static samples of casing damage are constructed as shown in the following table 3.

Table. 3. Machine learning samples based on static indicators

sysm	tslx	xch	xc hd	fsy hdb	nzsyhdb	ny hdb	sk ks	VSH	...
20	CD	P101	2.6	0.0	0.3	0.7	15	0.5	...
22	CD	P111	2.4	0.3	0.2	0.3	0	0.0	...
2	CD	P111	2.4	0.3	0.2	0.3	0	0.0	...
13	BX	P110	2.6	0.6	0.1	0.3	26	1.4	...
7	CD	P102	6.7	0.1	0.1	0.8	8	0.4	...
0	CD	P102	6.7	0.1	0.1	0.8	8	0.4	...
20	BX	P104	5.2	0.3	0.2	0.5	21	1.3	...
21	BX	P110	4.6	0.3	0.2	0.5	32	2.0	...
9	PL	P104	5.9	0.6	0.2	0.2	61	3.8	
20	BX	P110	2.6	0.6	0.0	0.4	53	1.1	
14	CD	P102	7.6	0.3	0.1	0.6	0	2.5	
18	BX	P109	4.6	0.1	0.5	0.5	31	1.4	
17	BX	P107	5.7	0.1	0.4	0.5	42	1.7	
...	...	...	...	...	...	...	...	...	...

B. Casing Damage Type Classification Model

Machine learning provides scientists with a set of tools for discovering new patterns, structures, and relationships in scientific datasets that are difficult to reveal through traditional techniques. The important theoretical basis of machine learning algorithms is classical statistics which is centered around the asymptotic theory when the number of samples approaches infinity. In the process of constructing casing damage prediction model for a single well, the samples used for modeling are often limited after filtering some data of blocks, wells, layers, etc. In particular, after all the influencing factors of casing damage are added, the machine learning faces problems such as small samples and high dimensions when it is adopted to analyze casing damage wells. Nonetheless, neural network algorithms often require more learning samples. When there are fewer samples, there are problems such as local extreme values and over-learning. Support vector machine transforms low-dimensional to high-dimensional through non-linear transformation on the principle of structural risk minimization, and then calculates the hyperplane to classify the data. However, when the dimension is too high, feature selection should be performed to meet the needs of SVM. Random forest is a classifier ensemble learning algorithm that does not rely on any model assumptions. It won't lose its action even in high-dimensional space and can achieve high prediction accuracy under any form of classification and regression. Also it is not prone to overfitting, enables MDA and MDI algorithms to evaluate the importance of features and has many other advantages. This paper mainly uses the random forest algorithm to establish classification and regression prediction models which are helpful in automatically finding rules and characteristics of layers where casing damages are about to occur from data of casing damage wells and wells without casing damages.

In order to analyze the influencing factors of casing damage at different horizons, a static sample set is constructed with the method described above based on static dataset of casing damage at Pubei Oilfield. Select "Small layer, small layer thickness(LT), siltstone thickness ratio(STR), argillaceous sandstone thickness ratio(ASTR), mudstone thickness ratio(MTR), perforation, perforation intervals, distance from the perforation top-bottom interface(MINPID), VSH, lithology, sand-mudstone interface distance(SMID)" and other indicators as input features, " types of casing damage" as a category label, and then a random forest algorithm is used to establish a recognition model of casing damages at different horizon. The identification results are shown in Table 4, which illustrates the differences of influencing factors in different formations. Take the three small layers of P101, P102, and P103 as examples: the sensitive factor of P101 is the "mudstone thickness ratio", the sensitive factor of P102 is the "sand-mud interface distance", and the sensitive factor of P103 is "small layer thickness".

The confusion matrix for the evaluation of model predictions is shown in Table 5. The average accuracy of casing damage identification models in different horizons is 84.2%. Among them, the prediction accuracy of casing deformation is 82.1%, and the recall accuracy is 95.8%; the prediction accuracy of casing fracture is 88.9%, the recall accuracy is 61.5%; the prediction accuracy of casing breakage/rupture is 100%, and the recall accuracy is 100%.

Table. 4. Recognition rules in different layers

<table>
<tr><th>Layer Number</th><th>Identification Rules</th></tr>
<tr><td>P101</td><td>XCH = P101
| MTR > 0.408
| | MTR > 0.597
| | | ASTR > 0.108: CD {BX=0, CD=3, PL=0}
| | | ASTR ≤ 0.108: BX {BX=7, CD=2, PL=0}
| | MTR ≤ 0.597: BX {BX=6, CD=0, PL=0}
| MTR ≤ 0.408: CD {BX=0, CD=5, PL=0}</td></tr>
<tr><td>P102</td><td>XCH = P102
| SMID > 0.227
| | MINPID > 18.190: CD {BX=0, CD=2, PL=0}
| | MINPID ≤ 18.190
| | | MINPID > 1.085: BX {BX=10, CD=0, PL=0}
| | | MINPID ≤ 1.085
| | | | MINPID > 0.770: CD {BX=0, CD=2, PL=0}
| | | | MINPID ≤ 0.770
| | | | | STR > 0.004
| | | | | | STR > 0.148: BX {BX=4, CD=0, PL=0}
| | | | | | STR ≤ 0.148: PL {BX=1, CD=0, PL=3}
| | | | | STR ≤ 0.004: BX {BX=3, CD=1, PL=0}
| SMID ≤ 0.227
| | ASTR > 0.172: BX {BX=2, CD=2, PL=0}
| | ASTR ≤ 0.172: CD {BX=0, CD=6, PL=0}</td></tr>
<tr><td>P103</td><td>XCH = P103
| LT > 6.350
| | ASTR > 0.179
| | | SMID > 1.305: CD {BX=1, CD=1, PL=0}
| | | SMID ≤ 1.305: BX {BX=5, CD=0, PL=0}
| | ASTR ≤ 0.179
| | | LT > 6.700: BX {BX=2, CD=1, PL=0}
| | | LT ≤ 6.700: CD {BX=0, CD=6, PL=0}
| LT ≤ 6.350
| | VSH > 0.947: BX {BX=14, CD=0, PL=0}
| | VSH ≤ 0.947
| | | ASTR > 0.135: BX {BX=8, CD=1, PL=0}
| | | ASTR ≤ 0.135: CD {BX=0, CD=2, PL=0}</td></tr>
</table>

Table. 5. Confusion matrix of casing damage identification model

	True：BX	True：PL	True：CD	Precision
Prediction：BX	23	0	5	82.14%
Prediction：PL	0	1	0	100%
Prediction：CD	1	0	8	88.89%
Recall	95.83%	100%	61.54%	-

IV. Block Casing Damage Risk Assessment

In order to evaluate the risk of casing damages in different blocks, the risk rank is evaluated according to the annual newly-added casing damages in blocks, as shown in the following table.

Table. 6. Risk ranks of casing damages in blocks

Levels	Casing damage risk level rules
High	Annual rate of newly-added casing damage ≥ 3%
Medium	1% ≤ Annual rate of newly-added casing damage ＜ 3%
Low	Annual rate of newly-added casing damage ＜ 1%

In order to assess risks of block casing damages, dynamic factors are designed. Oilfield development by water injection is a dynamic process. In different periods, due to different development plans, the casing damage rate changes dynamically with time and it is same case with other factors including formation pressure, differential pressure, water injection pressure, days of overpressure water injection, water injection intensity, completion of injection allocation, injection-production ratio, etc. The main indicators for evaluation of block casing damages are the pressure system and the injection-production relationship, including original formation pressure, average formation pressure, total pressure difference, injection-production pressure difference, water injection pressure difference, pressure difference among blocks, and cumulative injection-production ratio. Generally, the statistics of multiple wells in a block are used to represent the data characteristics of a single indicator, including the degree of concentration of data (mean, median, mode), degree of dispersion (standard deviation, coefficient of variation, quartile range), and distribution shape (skewness coefficient and kurtosis coefficient). In this case, one indictor will be split into eight indexes, which is easy to establish high-dimensional small sample data. Therefore, analysis of related row and similarity is performed on dynamic indicators of blocks.

Table. 7. The Samples for evaluation of casing damage risks in blocks

NF	JS	DQYL	ZYC	YJLY	ZCYC	LB	...
1994	95	7.959	-2.841	1.954	14.044	medium	...
1995	96	9.571	-1.229	2.69	15.419	low	...
1996	97	9.088	-1.712	2.145	15.421	medium	...
1997	102	9.726	-1.074	2.242	16.226	low	...
1998	109	8.943	-1.857	1.896	16.054	low	...
1999	113	9.803	-0.997	3.015	15.06	low	...
2000	114	9.619	-1.181	2.63	14.689	low	...
2001	114	9.625	-1.175	3.303	14.029	low	...
2002	115	9.76	-1.04	2.393	16.263	low	
2003	116	10.172	-0.628	3.005	15.092	medium	
2004	116	10.343	-0.457	2.527	16.016	medium	
2005	117	11.322	0.522	3.82	14.907	low	
...	...	...	...	...	...	...	...

We select the random forest algorithm to establish a risk identification model in blocks, as shown in the figure below. The average accuracy of the model is 80%. It can be seen that the main influencing factor of No.2 fault block of Pubei oilfield is the "water injection pressure difference".

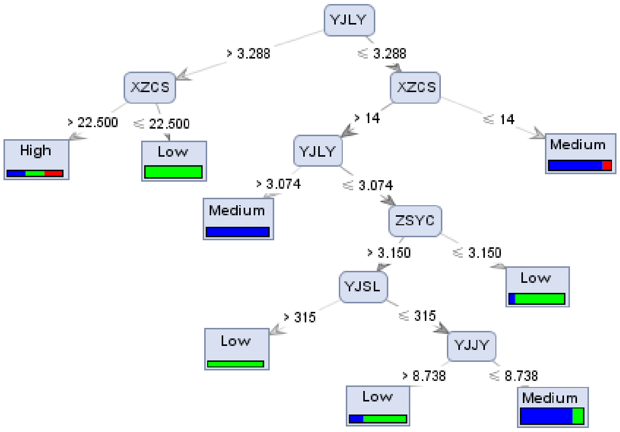

Fig. 3. Decision tree model for risk identification

V. Single well casing damage prediction

Oil or water wells with casing damages must be correctly judged and repaired in time to ensure the normal production of oilfields. Based on the different characteristics of casing damages shown by different data, this paper adopts machine learning algorithms to mine the key features hidden in them, so as to make single-well casing damage prediction more scientific, accurate and timely.

Sample generation. Extract the production data, measures data, perforation data, hierarchical data, casing damage and other data of the casing damage wells in the 4th fault block in Pubei Oilfield, and select oil pressure, casing pressure, water injection intensity, apparent water injectivity index , maximum allowable pressure difference, mainline pressure, daily water injection amount, monthly water injection amount, annual water injection amount, the number of days in production and other indicators as input features of the algorithm, and whether casing damages occur as the category label. A single well dynamic training sample set is constructed, with a total of 389 records, of which the records of casing damage is 65 and those of normal is 314.

Feature importance analysis. The random forest algorithm supports two methods, MDA and MDI, to effectively evaluate the importance of each feature in the modeling process, so as to determine the combination of features used in the modeling and exclude the effect of too many invalid features on the accuracy of the model. Table 9 shows the results of feature importance analysis of the casing damage prediction model for the injection wells in the fourth fault block in Pubei oilfield, indicating "oil pressure, maximum allowable differential pressure, mainline pressure, and water injection intensity" are the main influencing factors for casing damage of the injection wells in this fault block.

Table. 9. The MDA and MDI for the injection well factors

Factors	MeanDecreaseAccuracy	MeanDecreaseGini
oil pressure	33.07	32.17
mainline pressure	32.85	14.68
pressure difference	24.93	28.98
injectivity index	19.38	9.62
casing pressure	14.72	12.24
daily water injection	14.67	5.45
water injection intensity	13.78	5.23
water injection	12.79	5.17
production days	8.23	4.03
yearly water injection	8.09	3.48

Forecasting model. Random forest model is a classifier that uses multiple trees to conduct trainings and make predictions based on samples. Random forests consist of multiple classification and regression trees (CART) with each tree representing a decision tree. Each decision tree model quantitatively represents the rules of different parameters for casing damage identification and early warning.

Model evaluation. The confusion matrix of the early-warning model of single-well casing damages for the fourth fault block in Pubei oilfield is shown in Table 11, with an accuracy rate of approximately 95.6%.

Model application. The random forest model established in the above section is employed to make predictions for the other 18 wells without casing damages in the 4th fault block in Pubei oilfield. The prediction results are shown in Table 10. In order to verify the validity of the model, the operation records of 12 water wells in the first half of 2019 are extracted of which the prediction results of 11 wells are consistent with the operation results, with the coincidence rate about 91%.

The same forecasting method is used to extract the relevant data of the second fault block in Pubei oilfield to establish a prediction model of single-well casing damage for the block and make predictions for and carry out the verification of wells with casing damages in the block, with a coincidence rate about 75%. It can be seen that due to different number of wells and casing damage samples in different blocks, the accuracy and coincidence rates of the model are also different. The follow-up research will focus on the study on casing damage prediction methods based on small samples of blocks.

Table. 10. The prediction results of the fourth fault block

Well Number	Casing damage prediction probability	Casing damage prediction probability in the last 2 years	Final forecast results	Downhole operation verification
JH715	0.87	0.667	Casing damage	consistent
JH726	0.674	0.905	Casing damage	consistent
JH765	0.091	0.762	Normal	consistent
JH76F5	0.255	0.333	Normal	consistent
JH775	0.6	0.857	Normal	
JH7848	0.617	0.857	Normal	
JH7851	0.186	0.048	Normal	consistent
JH7950	0.723	0.952	Casing damage	inconsistent
JH8053	0.404	0.571	Normal	consistent
JH8052	1	0.952	Casing damage	consistent
JH8153	0.433	0.905	Normal	
JH8250	0.957	0.952	Casing damage	
JH8349	0.467	0.476	Normal	consistent
JH8351	0.352	0.905	Normal	
JH8353	0.623	0.667	Normal	consistent
JH8450	0.596	0.381	Normal	consistent
JH8751	0.023	0	Normal	
JH124	0.537	0.429	Normal	consistent

VI. Conclusion

On the basis of the data-driven concept, this paper establishes a set of methods for casing damage identification and prediction, including the construction of a big database of casing damage, dynamic and static sample generation, model construction, and model application. The established prediction model for single-well casing damage has a coincidence rate of 91% and 75% in the fourth and second fault block in Pubei oilfield, respectively, proving that the model has a good application value and provides a scientific basis and a clear direction for following prevention and management of casing damages. Compared with the traditional concept characterized by "reason-based measures, multiple remediation plans", the data-driven concept characterized by "data-driven decision-making, data-based governance" has truly made the automatic data service possible, and pushed the oilfield work model to transform from "digital mode" to "automatic mode " and "intelligent mode". However, petroleum data is a complex geophysical information system that spans a wide range of time and space. Even a simple business analysis involves many aspects of multiple discipline such as geology, exploration, logging, and development. The data quality, data integrity, random noise in data, and imbalance of the data set all pose a great challenge for the construction of machine learning data sets. Data is compared to petroleum in the new era. With the development of data science, a group of professional scientists in petroleum data will be born. Together with other professionals in petroleum, they will draw a grand blueprint for the past and present of petroleum data.

References

[1] LIU Ying. “Casing failure characteristics of the mature oilfields at home and abroad and suggestions of the failure prevention and control for Daqing Oilfield”, *Petroleum Geology & Oilfield Development in Daqing*, vol. 38, no.6, pp. 58-65, 2019.

[2] Ehsan Zabihi Naeini, Kenton Prindle. “Machine learning and learning from machines”, *The Leading Edge*, vol. 37, no.12, pp. 886-893, 2018.

[3] Bergen, Karianne J, Johnson, Paul A, et al. “*Machine learning for data-driven discovery in solid Earth geoscience*”. *Science*, vol. 363, no.6433, pp. 1299-1305, 2019.

[4] Zhou Xiangguang, Li Dawei. “Prediction of casing failure by gradient boosting decision tree algorithm”, *Journal of Computer Applications*, vol. 38, no.S2, pp. 144-147, 2018.

[5] Zhang Xu, Wang Lu, Meng Fanshun, Zheng Zhichao. “Bayesian neural network approach to casing damage forecasting”, *Progress in Geophysics*, vol. 33, no.3, pp. 1319-1324, 2018.

[6] Huang Jun, Meng Fanshun, Zhang Xu, Yang Guanyu. “Application of genetic neural network based on PCA in prediction of casing damage”, *Journal of Xi'an Shiyou University(Natural Science Edition)*, vol. 33, no.6, pp. 84-89, 2018.

[7] Yu Guijie, Zhao Chong, Chi Jianwei, Zhang Jiaxing. “Fatigue life prediction of coiled tubings based on artificial neural network”, *Journal of China University of Petroleum(Edition of Natural Science)*, vol. 42, no.3, pp. 131-136, 2018.

[8] Jiang Xueyan, Zhang Shujuan, Wang Zhiguo, Liu Hailong, Zhao Chunyu, Wang Hejun. “Evaluating method of geological factor risks of the casing damage based on certainty factors”, *Petroleum Geology & Oilfield Development in Daqing*, vol. 35, no.6, pp. 104-108, 2016.

.

Yanhong Zhao was born on Mar. 3, 1986. She received the PhD degree in the Petroleum Engineering and Data Mining Lab of College of Geophysics and Information Engineering at China University of Petroleum-Beijing. Her current research interests include machine learning and knowledge management in the petroleum field.

Hanqiao Jiang was born on Aug.6, 1957**.** He is a professor and doctoral supervisor of China University of Petroleum (Beijing). He graduated from the Development Department of East China Institute of petroleum in 1982. In 1997, he worked as a senior visiting scholar in the EOR center of the University of Wyoming. In 1997, he was promoted to a professor. In 2000, he was employed as a doctoral supervisor and academic leader of oil and gas field development engineering.

Hongqi Li was born on Jan. 4, 1960**.** He is a professor at College of Geophysics and Information Engineering in China University of Petroleum-Beijing. He received the Ph.D. degree with the major of Solid Earth Geophysics at Institute of Geophysics, Chinese Academy of Sciences in 1998. He was a visiting scholar at Texas A&M University in 1989 and Connecticut University in 1999, respectively. His current research interests include knowledge discovery and data mining, intelligent construction of petroleum and logging reservoir evaluation.