

Analysis of Open Well Datasets

D.O. Makienko^{1,A}, I.V. Safonov^{2,B}

^A LLC TCS, Moscow Research Center, Moscow, Russia

^B National Research Nuclear University MEPhI, Moscow, Russia

¹ ORCID: 0000-0001-7341-6128, makienko-dasha@mail.ru

² ORCID: 0000-0002-8270-943X, ilia.safonov@gmail.com

Abstract

Recently, the number of studies devoted to the use of machine learning methods in geophysics has been increasing significantly. Examples of such investigations include the prediction of rock properties and separation of rock types according to quantitative characteristics. Annotated datasets are required to build and evaluate the quality of machine learning based models. This paper analyzes open labeled well datasets and related research. We consider data containing well logs, rock images, laboratory results, labeled zonation by lithotypes. Methods for visualizing well data are presented. We provide recommendations for oil and gas companies on the preferable format for making well data publicly available

Keywords: Well logs, rock images, open datasets, machine learning.

1. Introduction

In the last decade, progress in many fields has been driven by the widespread application of machine learning (ML) methods. Successfully solving tasks using ML methods is typically associated with the availability of a large, representative set of labeled data. However, researchers often encounter situations, where diverse labeled data are insufficient to create models with high generalization ability. Data augmentation and the generation of artificial data can significantly improve the quality of solutions in some cases [1]. The use of open datasets allows researchers to test the applicability and evaluate the generalization ability of existing ML models and sometimes improve them.

In the development of oil and gas fields, geophysical well logging is performed [2,3]. During well logging, sensors are lowered into the well, which measure rock properties (electrical, radioactive, acoustic, and others) at certain depths and times. The measurement results are presented as data arrays referenced to the depth of the well. When drilling, rock samples (core) are extracted from the well, photographed, and specimens are cut from the core for laboratory analysis of various rock properties, such as porosity and permeability.

This research object represents a typical example of multidimensional and heterogeneous data. The tasks addressed using well data are a specific case of multidimensional data analysis and visualization. The approaches utilized in the oil and gas industry can be adapted to other domains, such as those dealing with time-series data.

The oil and gas industry has established a data visualization approach as illustrated in Figure 1. The first column contains the depth scale. The second column displays core photographs. The remaining columns show well logs for various sensor types: DENS – bulk density, DTC – compressional wave travel time, DTS – shear wave travel time, GR – gamma ray, NEUT – neutron porosity, REF – photoelectric factor, and RT – resistivity. Working with well data often involves dealing with missing information at certain depths. In Figure 1, core photographs are missing for some depth intervals, but the well logs are complete. However, such gaps are frequently encountered in practice, and the task of filling incomplete data is highly relevant [4].

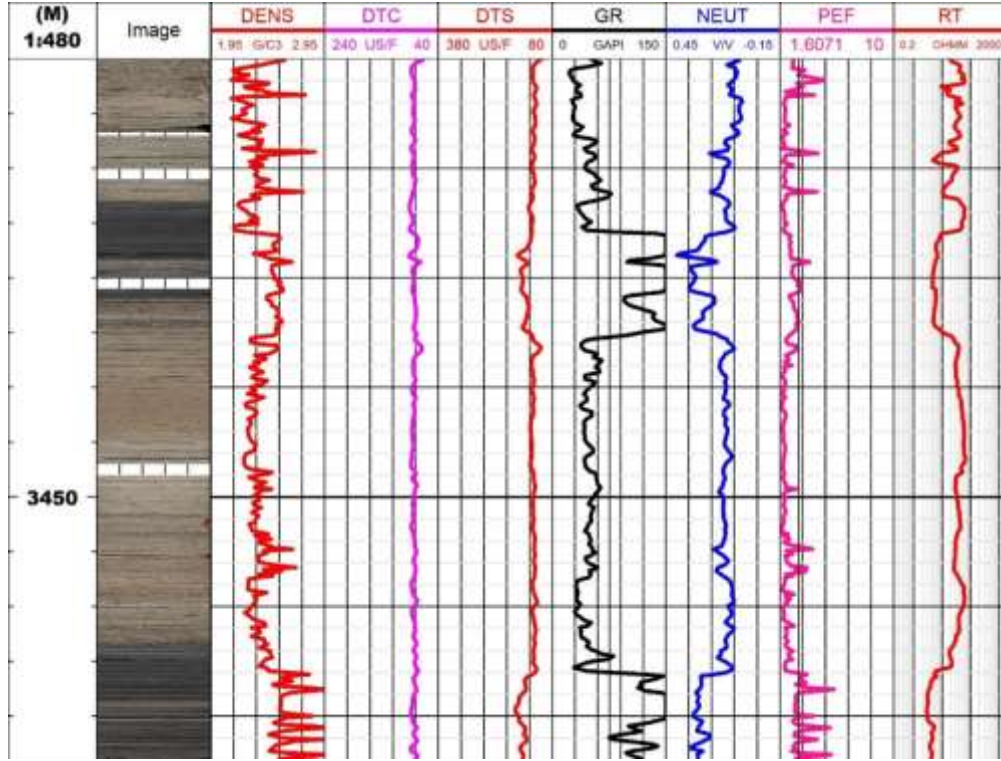


Figure 1 - Example of well data visualization

Based on NOPIMS well data by Geoscience Australia which is © Commonwealth of Australia and is provided under a Creative Commons Attribution 4.0 International License and is subject to the disclaimer of warranties in section 5 of that license.

Over the past decades, a vast amount of well data has been accumulated. However, these data belong to the companies and are generally classified as confidential. Additionally, several countries, including the Russian Federation [5], have restrictions on the export of geological information, which prevents the disclosure of well log data. For these reasons, even the few existing open well datasets can significantly aid in the development of methods for their processing and analysis.

This paper presents a comparative analysis of well data from five datasets used in competitions for applying machine learning (ML) to geophysical data analysis. It discusses tasks that have been or can be solved using these datasets, demonstrates a method for visualizing well logs, and shows typical quality issues of well data. Additionally, the paper provides an overview of several online resources offering open-access well data. The article significantly expands upon the review conducted by the authors in [6], particularly by providing a detailed analysis of class imbalance and missing data, demonstrating outliers in competition datasets, and offering recommendations for oil and gas companies on the preferred format for making well data publicly available.

2. Datasets overview

Table 1 provides information about the datasets used in five competitions for analyzing well logs. The FORCE 2020 Machine Learning Competition is referred to as FORCE-2020 in the table and further in the article. This competition addressed two tasks: lithofacies classification and fault mapping on seismic data. We focus only on the first task. The dataset includes one training set and two test sets. The 2016-ml-contest also focused on lithofacies prediction and contains both training and test datasets. Competitions organized by the SPWLA Petrophysical Data-Driven Analytics Special Interest Group are labeled as 2020-SPWLA, 2021-SPWLA, and 2023-SPWLA according to the year of the competition. The 2020-SPWLA competition involved predicting missing well logs. The 2021-SPWLA competition aimed at

predicting rock properties. The 2023-SPWLA competition focused on matching the depth of well logs. For 2020-SPWLA and 2021-SPWLA, both training and test datasets are available. For 2023-SPWLA, a training dataset and input data for testing predictive models are accessible.

Table 1 – Datasets from well log analysis competitions

Dataset	FORCE-2020 ^{1,2} [7]			2016-ml-contest ^{4,5}		2020-SPWLA ⁶		2021-SPWLA ⁶		2023-SPWLA ⁶	
	Train	Test 1	Test 2	Train	Test	Train	Test	Train	Test	Train	Test
Number of wells	98	10	10	10	2	3	4	9	4	9	3
Number of rows in tables	1170511	136786	122397	3232, 4149	809	30143	11088	318967	11275	69304	19038
Count of data types	29	28	29	11		9		17		5	11
Size in MB	267	31	29	0.2	0.05	1.9	0.7	37	1	2.4	0.9
Task	Classification of 12 rock classes			Classification of 9 rock classes		Regression for predicting well logs		Regression for predicting rock characteristics		Depth matching	
License	CC-BY 4.0 ² , NLOD 2.0 ³			CC0 1.0 ⁵		CC BY-NC-SA, Equinor Open Data License ⁷		CC BY-NC-SA, Equinor Open Data License ⁷		not specified	

¹ <https://github.com/bolgebrygg/Force-2020-Machine-Learning-competition>

² <https://doi.org/10.5281/zenodo.4351156>

³ Contains data under the Norwegian license for Open Government Data (NLOD) distributed by Norwegian government.

⁴ <https://github.com/seg/2016-ml-contest>

⁵ <https://www.kaggle.com/datasets/imeintanis/well-log-facies-dataset>

⁶ <https://github.com/pddasig>

⁷ Contains data under the Equinor Open Data License distributed by Equinor and the former Volve license partners.

The datasets presented in Table 1 pertain to well measurements, and one of the characteristics of these datasets is the number of wells. For all competitions, the training and test datasets use data from different wells but from the same field. Each dataset is a table stored in a CSV file. The data files are characterized by the number of rows in the tables, the number of columns with different types of measurements, and the file size in megabytes. The number of rows refers to the number of measurements at different well depths. Data from different wells are either separated into different files or combined into a single table. The well's identification number or name may be included in the combined table.

For the 2016-ml-contest, there were no missing rows in the training and test datasets; however, there was a file that included the training set and an additional 917 rows for which one well log was missing. Therefore, the training set from the 2016-ml-contest is characterized by two values for the number of rows. In the FORCE-2020 competition, the first test set has one less data type than the training set and the second test set. The test dataset for the 2023-SPWLA competition includes six additional fields for prediction results.

Figure 2 illustrates data from four wells in the 2021-SPWLA test dataset. The well logs were plotted using the matplotlib package (<https://matplotlib.org>) for the Python programming language. To visualize missing values in well data, the missingno package (<https://github.com/ResidentMario/missingno>) is convenient. The numbers 1 and 11275 on the left side of the plot correspond to the first and last row numbers in the table. The leftmost column indicates the well number. The gray plot on the right side shows the number of values in each row of the table. This graph displays the minimum and maximum number of com-

plete data points for the rows in the table. White areas in the columns with well logs correspond to missing values.

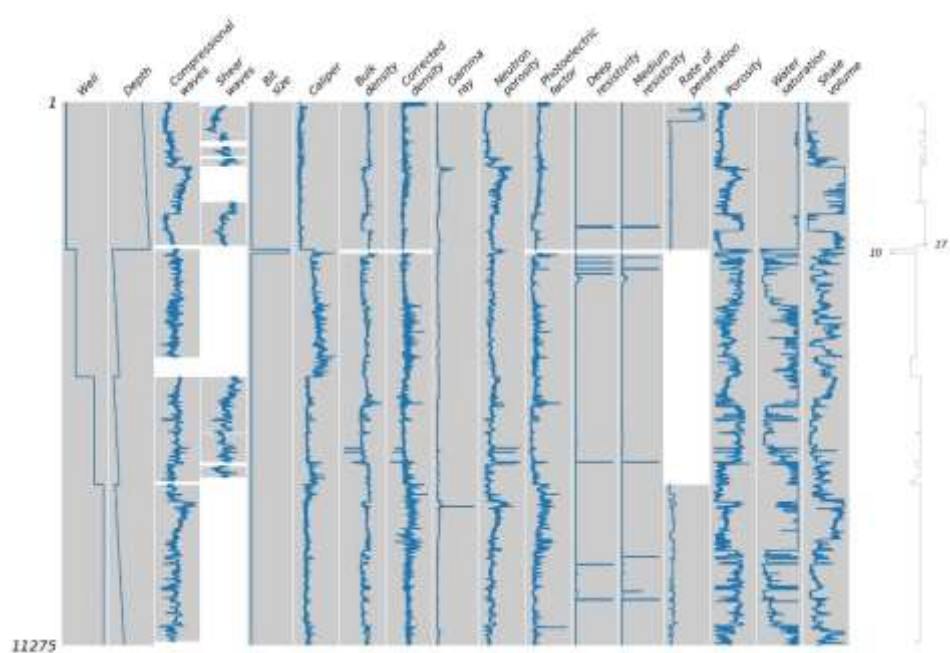


Figure 2 - Visualization of the 2021-SPWLA test dataset

The missingno library allows for demonstrating the correlation between missing values for pairs of data columns. Figure 3 shows a correlation matrix of missing values for the columns in the 2021-SPWLA test dataset. The axes list the types of well logs. The color and numbers indicate the correlation value for pairs of well logs. For insignificant correlations, the numerical value is not displayed. Labels "<1" or ">-1" correspond to cases, where the correlation value is close to 1 or -1 respectively. Analyzing the correlations of missing values can be useful when deciding whether to fill or delete rows with missing values. For instance, one can fill in the gaps simultaneously in several well logs with high correlation. On the other hand, weak correlation between pairs of well logs indicates which logs will lose data when rows with missing values are removed.

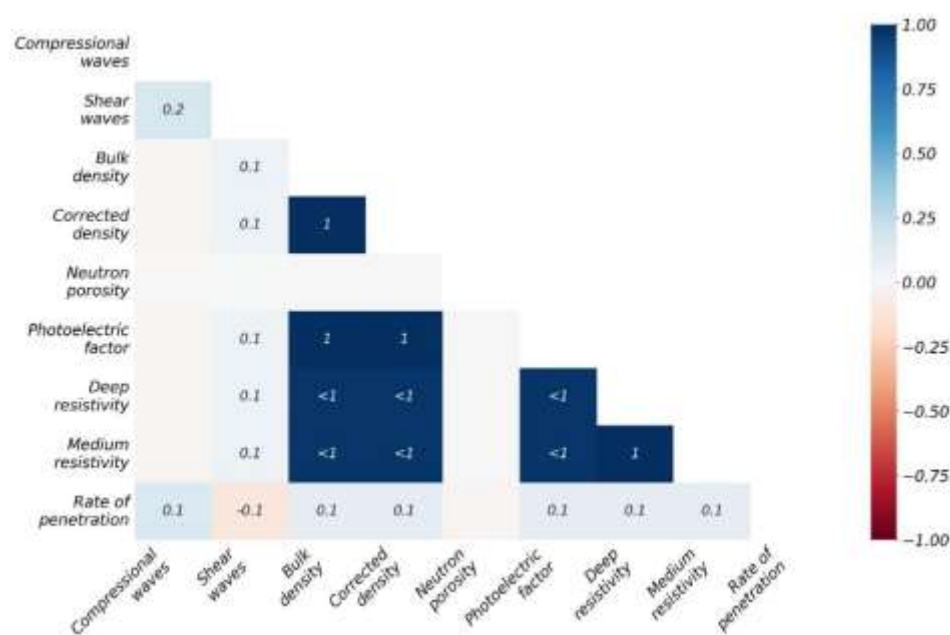


Figure 3 – Missing value correlation matrix for well logs from the 2021-SPWLA test dataset

In the FORCE-2020 competition, the task was to classify twelve lithofacies classes. Many well datasets have class imbalances due to varying frequencies of different rock types occurrences. The class distribution in the FORCE-2020 training set is shown in Figure 4. A similar class distribution is observed in the test datasets. The 2016-ml-contest involved the classification of nine rock classes. The 2020-SPWLA dataset is intended to predict acoustic well logs. The 2021-SPWLA competition involved a regression task to predict shale volume, porosity, and water saturation. The 2023-SPWLA competition focused on matching the depths between well logs.

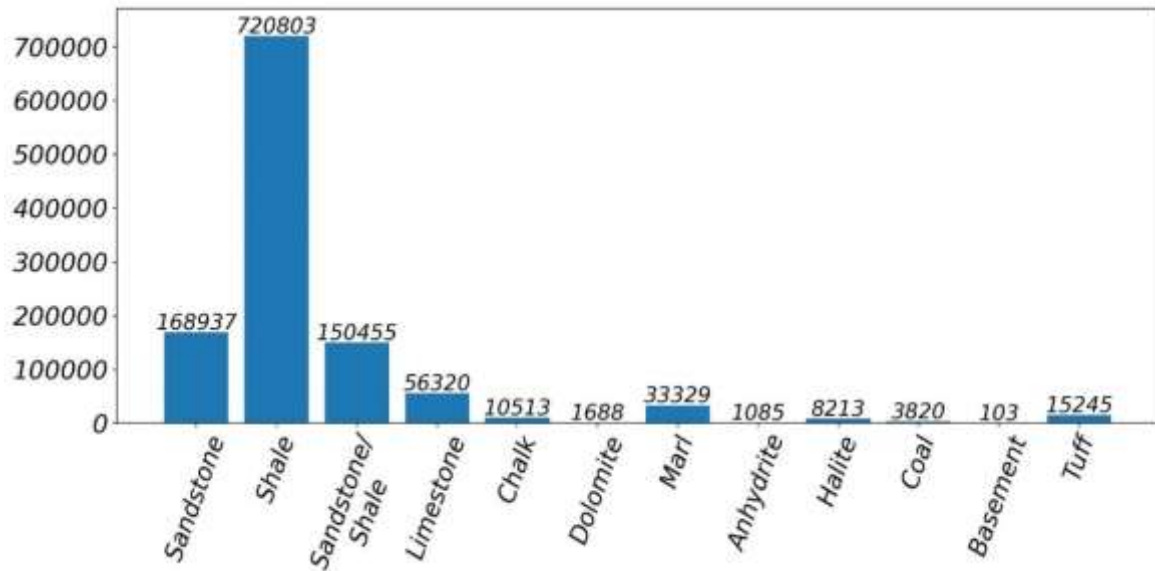


Figure 4 - Distribution of lithofacies classes in the FORCE-2020 training dataset

Table 2 characterizes class imbalance in the FORCE-2020 and 2016-ml-contest datasets. Class imbalance affects the quality of the classification model; therefore, it is recommended to consider it during training [8]. For each dataset, the number of classes, the largest and smallest class sizes, and the ratio of the smallest to the largest class size are indicated. The ratio of the smallest to the largest class size ranges from 0.0001 for the highest imbalance to 0.15 for the case where the imbalance is not as significant.

Table 2 – Class imbalance in well datasets

Dataset		Number of classes	Largest class size	Smallest class size	Ratio of the smallest class size to the largest
FORCE-2020	Train	12	720803	103	0.0001
	Test 1	10	83875	416	0.0050
	Test 2	11	71827	244	0.0034
2016-ml-contest	Train	9	940	141	0.1500
	Test	10	166	6	0.0361

Table 3 lists the types of data present in the competition datasets. The well logs are categorized into several groups: gamma ray, electrical resistance, neutron porosity and bulk density, acoustic measurements (travel time), photoelectric factor, and spontaneous potential. There is also a group with interpretation results. For some data types, clarifying comments are provided in parentheses. Explanations of the physical meaning of well log data and recommendations for their use in geophysical interpretation can be found in references [2,3]. In addition to the listed types of data, the datasets include metadata containing drilling information.

Table 3 – Types of well logs in competition datasets

Datasets	FORCE-2020	2016-ml-contest	2020-SPWLA	2021-SPWLA	2023-SPWLA
Data type	1429694 ¹	4958	41231	330242	88342
Gamma ray	100.00 ² 4.93 (spectral)	100.00	99.38	99.06	100.00
Electrical resistance	48.64 (shallow) 96.54 (medium) 99.22 (deep) 14.17 (micro) 24.34 (flushed zone)	100.00	99.07 (medium, deep)	90.83 (medium) 90.76 (deep)	100.00 (deep)
Neutron porosity and bulk density	67.57 (neutron porosity) 86.87 (bulk density) 84.75 (density correction)	100.00 (neutron-density porosity difference) 100.00 (average neutron-density porosity)	98.22 (neutron porosity) 98.35 (bulk density)	32.34 (bulk density) 30.37 (density correction) 32.31 (neutron porosity)	100.00 (neutron porosity) 100.00 (bulk density)
Acoustic measurements	94.00 (compressional) 20.33 (shear)		90.17 (compressional) 88.20 (shear)	24.38 (compressional) 14.46 (shear)	
Photoelectric factor	61.95	81.50	98.35	30.20	
Self (spontaneous) potential	68.35				
Interpretation	100.00 (lithofacies) 90.42 (confidence)	100.00 (lithofacies)		17.74 (porosity) 17.74 (water saturation) 17.07 (shale volume)	
Percentage of rows that have no missing values	0.00	81.50	77.07	3.36	100.00

¹ total number of rows in the dataset

² percentage of non-missing rows in a column of a given type

The first row of Table 3 lists the dataset names and the total number of rows for the training and test data combined. The table cells indicate the percentage of non-missing rows in the columns of that data type. Values of -999, -999.25, -999.9, and -9999 in the datasets correspond to missing values. Other values indicating the absence of data may be identified during the outlier detection stage. The last row of Table 3 shows the percentage of rows in the dataset that have no missing values.

Figure 5 shows box plots for the well logs from the FORCE-2020 training dataset. The blue rectangles represent the range of values from the 0.25 to the 0.75 quantile distribution. The horizontal line within the rectangle corresponds to the median, while the white circle represents the mean. The difference between the 0.75 and 0.25 quantiles is the interquartile range. Values greater than the 0.75 quantile or less than the 0.25 quantile by 1.5 times the interquartile range are shown as lines extending from the rectangle. Sometimes values outside these lines can be considered as outliers, but typically additional data analysis is required to determine outliers confidently. Reference [9] provides an example where initially several gamma ray values are identified as outliers, but further analysis reveals that they may correspond to genuine geological features.

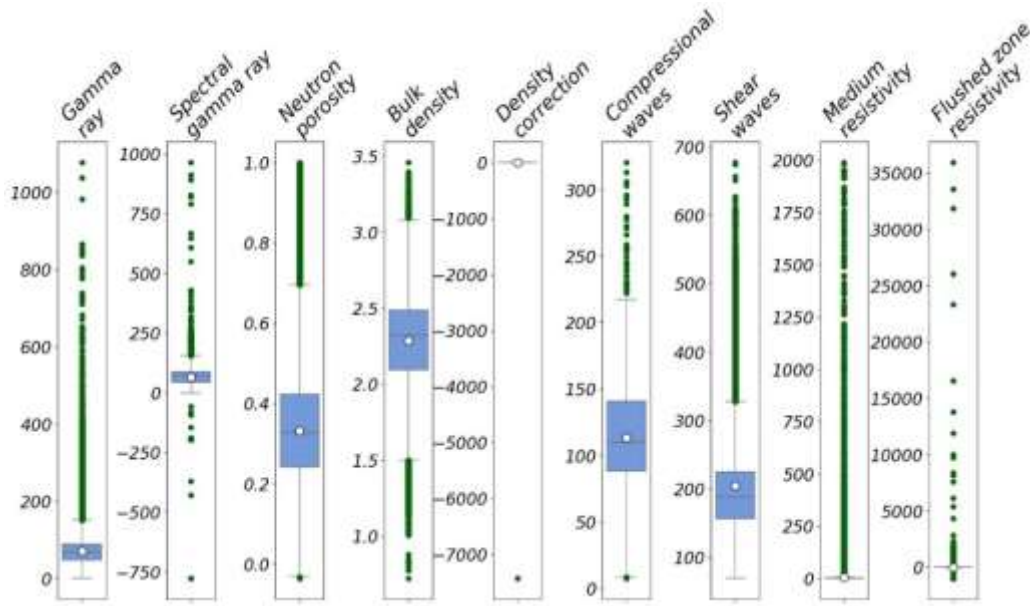


Figure 5 – Box plots for well logs from the FORCE-2020 training dataset

Methods that operate simultaneously with multiple log types can be used to detect outliers. Figure 6 illustrates outliers in the 2020-SPWLA test dataset, identified using the "isolation forest" method [10]. The outliers are shown as red dots. The implementation of the "isolation forest" method from the scikit-learn library (<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html>) is used. Outliers are detected for all columns.

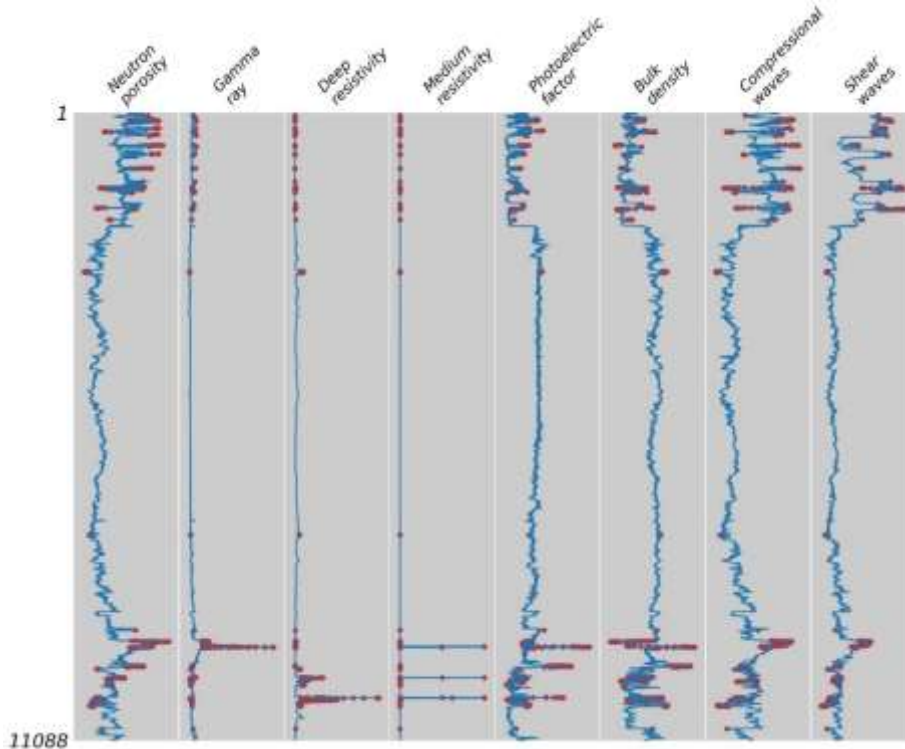


Figure 6 – Outliers in well logs from the 2020-SPWLA test dataset determined by the "isolated forest" method

There are several online resources containing open-access well data. Unlike competition datasets for well log analysis, significant efforts are required to preprocess these data, par-

ticularly to compile them into depth-referenced tables. Table 4 provides general information about open-access well data available from the following five portals:

- WAPIMS (Western Australian Petroleum and Geothermal Information Management System) provides drilling information for wells in Australia.
- NOPIMS (National Offshore Petroleum Information Management System) contains data on offshore wells in Australia.
- BGS (British Geological Survey) offers open-access geological data from the United Kingdom.
- Volve is a dataset for a field in the North Sea, developed by Equinor company.
- NLOG provides information on the energy and mineral resources of the Netherlands.

Table 4 – Open well data resources

Resource	Well examples	Well logs	Well log interpretation	Images	Lithological zonation	Routine core analysis	License
WAPIMS ¹	GSWA Harvey1, DMP Harvey 2, DMP Harvey 3, DMP Harvey 4	yes	-	yes	yes	yes	CC-BY 4.0
NOPIMS ²	Satyr 5, Dorado 2, Barossa 4	yes	-	yes	-	yes	CC-BY 4.0
	Dorado 3, BDC 4 06 P	yes	-	yes	-	-	
BGS ³	204/19-3A, 204/20- 2, 10/01-A25, 106/20- 1	-	-	yes	-	-	Open Government License v3.0
Volve ⁴	15/9-F-1, 15/9-F-1A, 15/9-F-1B	yes	yes	-	-	-	Equinor Open Data License
NLOG ⁵	F13-01	yes	-	-	yes	-	Users are permitted to copy, to download and to disclose in any way, to distribute or to simplify the information provided on this website without the prior written permission of NLOG.NL or the lawful consent of the entitled party. Users are also permitted to copy, duplicate, process or edit the information and/or layout, provided NLOG.NL is quoted as the source. ⁶
	K05-02	yes	-	-	-	yes	
	ANNERVEEN-06	yes	-	yes	-	yes	

¹ <https://wapims.dmp.wa.gov.au/WAPIMS>,

² <https://nopims.dmp.wa.gov.au/Nopims>,

³ <https://webapps.bgs.ac.uk/data/offshoreWells>,

⁴ <https://www.equinor.com/energy/volve-data-sharing>,

⁵ <https://www.nlog.nl/en>,

⁶ <https://www.nlog.nl/en/disclaimer>

Table 4 includes information on the availability of well logs, their interpretations, images, lithological zonation, and results of routine core analysis. Dashes in the table cells indicate

the absence of such data or the inability to access them. The Volve dataset contains well logs for twenty-four wells and is used in SPWLA competitions. The WAPIMS, NOPIMS, BGS, and NLOG websites provide information on a large number of wells, and the types of data available can vary. The names of the specific wells considered are listed in the table. On the BGS website, data from the section on offshore hydrocarbon wells are considered.

Often well logs are in reports with images of logs, for example, in PDF format. Since extracting data from well log images is a non-trivial task, such data formats were not considered in our review. Table 4 includes information available in numerical form, stored in industry-specific formats such as LAS or DLIS. Python packages, such as lasio (<https://lasio.readthedocs.io/en/latest/>) and dlisio (<https://dlisio.readthedocs.io/en/latest/index.html>), can be used to work with these formats.

Lithological zonation refers to the description of rock types for depth intervals. Lithological zonation can be stored as tables, text, or schematic images. It is described with varying levels of detail. Routine core analysis refers to the results of laboratory studies on samples extracted from the core at certain depths. Routine core analysis determines rock characteristics, such as porosity and permeability.

Core images are obtained by photographing the rocks under natural and ultraviolet light, as well as using computed tomography [11,12]. Examples of rock images from various resources are shown in Figure 7. The extracted core is usually placed in boxes labeled with the corresponding depth intervals. A single photograph typically includes images of several boxes. Core photographs may feature circular holes resulting from the extraction of samples for further laboratory analysis. Wooden, cardboard, or other inserts are placed in the gaps where the rock is missing.

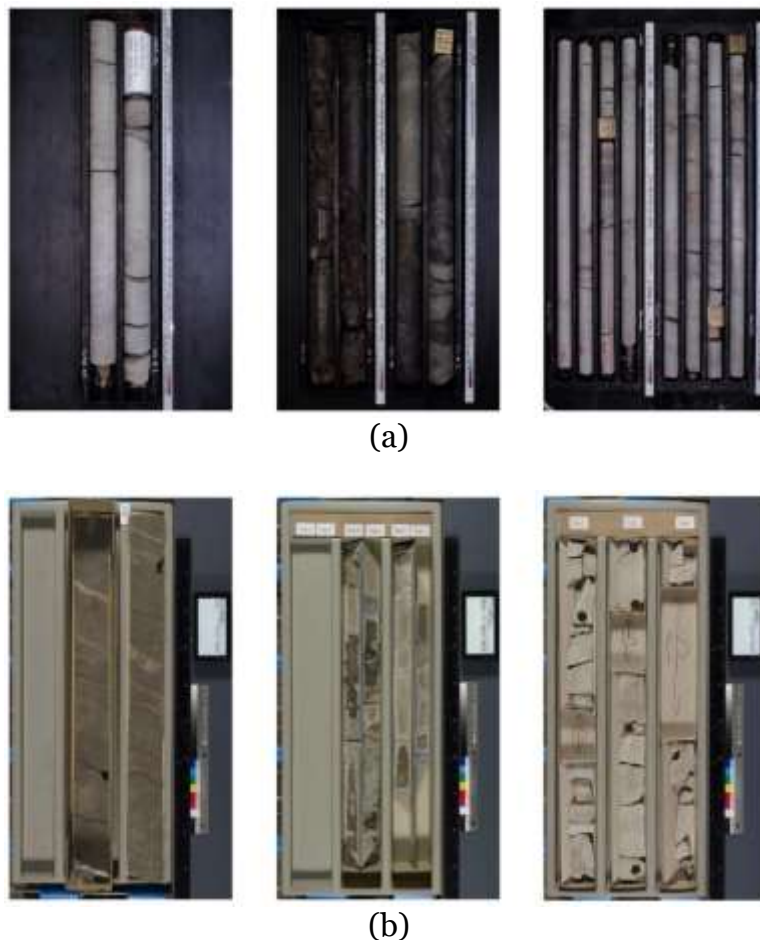


Figure 7 – Examples of rock images: (a) WAPIMS¹; (b) BGS²

¹ © State of Western Australia (Department of Energy, Mines, Industry Regulation and Safety)

² Contains British Geological Survey materials © UKRI [2024]

3. Review of tasks addressed using well data

Table 5 lists the tasks addressed using the well datasets under consideration. The second column enumerates the competition datasets and Internet resources. When using data from these online resources in publications, examples of the wells used are typically provided. The third column lists the studies conducted on these data, in the form of publications or solutions proposed by competition participants. Competition solutions are available as programs or brief reviews.

Table 5 – Tasks addressed using well data

Task	Data source	Research
Rock classification	FORCE-2020	Competition solutions, [13,14]
	2016-ml-contest	Competition solutions, [15,16,17]
	WAPIMS	[18]
	BGS	[19]
	NLOG	[20,21]
	NOPIMS	[22]
Predicting rock characteristics	2021-SPWLA / Volve	Competition solutions
	WAPIMS	[23]
	Volve	[9]
Depth matching	2023-SPWLA	Competition solutions
Filling gaps in well logs	FORCE-2020	[24]
	Volve	[9,24]
	NLOG	[25]
Well log prediction	2020-SPWLA / Volve	Competition solutions, [26,27]
	NOPIMS	[28]
Identification of undisturbed rock fragments in images	BGS	[29]
Finding correlations between geological sections of different wells	FORCE-2020	[30]

Rock classification was performed using both well logs and images. Lithological zonation was present for the FORCE-2020 and 2016-ml-contest datasets, as well as wells obtained from the WAPIMS and NLOG resources in studies [18,20]. For wells from the BGS, NLOG, and NOPIMS resources used in studies [19,21,22], the authors obtained lithological zonation manually or semi-automatically.

The regression task for predicting rock properties was addressed using three datasets. The 2021-SPWLA competition data were used to predict shale volume, porosity, and water saturation. The WAPIMS resource well logs were used in study [23] to predict porosity and permeability. Study [9] evaluated the impact of well logs selection on the quality of porosity prediction using the Volve dataset.

Well log predictions were made using the 2020-SPWLA and NOPIMS datasets, with acoustic logs being predicted for both. Filling gaps in well logs was performed using the

FORCE-2020, Volve, and NLOG datasets. This task is similar to predicting rock properties and well logs, where the target variable is continuous. However, in general, not only regression models but also other techniques, such as filling in missing values with mean values, can be used to fill missing values.

Different approaches are taken for data preprocessing in these studies. For well logs, preprocessing may include:

- Selecting the most significant well logs based on their physical understanding or using specialized algorithms.
- Augmenting the set of well logs with new logs obtained by transforming the originals.
- Filling or removing missing values.
- Detecting and removing outliers.
- Matching depths between different well logs and rock images.

For rock images, preprocessing includes isolating the rock from the background of the boxes, matching illumination and color balancing, and removing various types of defects. In study [29], undisturbed rock fragments are identified in images from the BGS dataset.

In addition, the FORCE-2020 dataset was used to solve the task of finding correlations between geological sections of different wells [30].

Typically, studies are conducted within a single dataset containing information about wells from one or several nearby fields. The quality of models is assessed on wells or well sections not involved in training. Although there are challenges in building models even within a single dataset, it would be interesting to attempt to build models applicable to different datasets.

4. Recommendations on the format for making well data publicly available

For the development of well data processing methods using ML, it is important to engage a broad scientific community, not limited to specialists within oil and gas companies. A large volume of data must be made available in open access. This paper considers two types of sources for well data: competition datasets and online resources. Competition datasets are conveniently prepared to enable ML engineers without domain knowledge of the oil and gas industry to start working with them. Using datasets from online resources requires deeper immersion in the specifics of the industry and numerous preprocessing steps, such as extracting numerical measurements from well log images in PDF files and matching measurements by depth.

Therefore, it is preferable to provide data similar to competition datasets, specifically:

- Distribute data in CSV format.
- Provide depth-matched data.
- Include metadata about the recording equipment and measurement conditions.

Photographing rock samples is done under various lighting conditions and with different devices. It is recommended to place special color charts in the field of view during photography. The presence of such charts allows for color correction of the images.

In countries with restrictions on the export of geological information, it is necessary to provide well information in an anonymized form. Approaches of data transformations for anonymization include, but not limited:

- Removing or assigning conditional names to fields and wells.
- Removing or assigning conditional depths to measurements.
- Grouping and mixing data from different wells and fields based on certain assumptions.

5. Conclusion

Progress in the field of machine learning is largely driven by the quantity and quality of available data. Providing open access to labeled well data promotes the further development

of ML models applied in geophysics. This paper presents a review and analysis of existing open well datasets. It considers five datasets used in competitions for applying ML to solve geophysical problems and five online resources providing various types of well data. We present typical approaches to well data visualization and list open-source libraries for such visualization. We describe the challenges faced by ML engineers when working with well data and provide the recommendations for oil and gas companies on preferred format the data.

The following tasks were addressed using the well datasets considered:

- Classification of rock types based on well logs and images.
- Prediction of rock properties from well logs.
- Prediction of acoustic well logs.
- Correction of well logs, which included filling gaps and matching depths.
- Finding correlations between geological sections of different wells based on well logs.
- Identifying undisturbed rock fragments in images.

The following tasks can be expected to be solved in the future using these datasets:

- Building general models for multiple fields.
- For well logs: creating domain-specific data filtering methods, such as cleaning logs from outliers; developing methods to correct shifts in well logs caused by measurements taken with different instruments under varying conditions.
- For images: identifying and correcting defects and determining various types of inclusions within rocks.
- Correcting class imbalance of rock types when building predictive models.

References

1. Deep neural networks for ring artifacts segmentation and corrections in fragments of CT images / A. Kornilov, I. Safonov, I. Reimers, I. Yakimchuk // 28th Conference of Open Innovation Association (FRUCT) (Moscow, 25-29 January 2021). IEEE, 2021. P. 181-193. DOI: <https://doi.org/10.23919/FRUCT50888.2021.9347587>.
2. Ellanskii M. M. Izvlechenie iz skvazhinnykh dannykh informatsii dlya resheniya poiskovo-razvedochnykh zadach neftegazovoi geologii [Extracting information from well data to solve exploration problems in oil and gas geology] Moscow: Gubkin University Press, 2000. 80 p. [in Russian]
3. Koskov V. N., Koskov B. V. Geofizicheskie issledovaniya skvazhin i interpretatsiya dannykh GIS [Geophysical surveys of wells and interpretation of well log data] Perm: Publishing house of Perm State Technical University, 2007. 317 p. [in Russian]
4. McDonald A. Impact of Missing Data on Petrophysical Regression-Based Machine Learning Model Performance // The SPWLA 63rd Annual Logging Symposium (Stavanger, Norway, June 2022). OnePetro, 2022. <https://doi.org/10.30632/SPWLA-2022-0125>
5. Klopov A. V. Osobennosti eksporta geologicheskoi informatsii v tsifrovuyu epokhu [Features of exporting geological information in the digital era] [Online] // Molodoi uchenyi. 2018. №. 41. P. 7-9. URL: <https://moluch.ru/archive/227/53043/> (accessed on 03.10.2023). [in Russian]
6. Makienko D. O., Safonov I. V. Obzor otkrytykh naborov skvazhinnykh dannykh [Overview of open well datasets] // 33rd International Conference on Computer Graphics and Computer Vision (Moscow, 19-21 September, 2023): conference proceedings / Graphicon Conference on Computer Graphics and Vision. Moscow: Institute of Applied Mathematics named after. M. V. Keldysh, Russian Academy of Sciences, 2023. V. 33. P. 710-720. [in Russian] <https://doi.org/10.20948/graphicon-2023-710-720>
7. Bormann, P., Aursand, P., Dilib, F., Manral, S., Dischington, P. (2020). FORCE 2020 Well well log and lithofacies dataset for machine learning competition [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.4351156>
8. Makienko D., Seleznev I., Safonov I. The effect of the imbalanced training dataset on the quality of classification of lithotypes via whole core photos // VI International Conference

on Information Technology and Nanotechnology (ITNT-2020) (Samara, 26-29 May 2020). CEUR Workshop Proceedings. 2020. T. 2667. P. 132-136.

9. McDonald A. Data Quality Considerations for Petrophysical Machine-Learning Models // Petrophysics. 2021. V. 62. №. 06. P. 585-613. DOI: <https://doi.org/10.30632/PJV62N6-2021a1>.

10. Liu F. T., Ting K. M., Zhou Z. H. Isolation Forest // The 8th IEEE International Conference on Data Mining (ICDM 2008) (Pisa, 15-19 December 2008) IEEE, 2008. P. 413-422. <https://doi.org/10.1109/ICDM.2008.17>

11. Visualization of quality of 3D tomographic images in construction of digital rock model / A.S. Kornilov, I.A. Reimers, I.V. Safonov, I.V. Yakimchuk // Scientific Visualization, 2020. V. 12. № 1. P. 70-82. <https://doi.org/10.26583/sv.12.1.06>

12. Kornilov A., Safonov I., Yakimchuk I. Blind quality assessment for slice of microtomographic image // The 24th Conference of Open Innovations Association (FRUCT) (Moscow, Russia, 8-12 April 2019). IEEE, 2019. P. 170-178. <https://doi.org/10.23919/FRUCT.2019.8711938>

13. Feng R. Uncertainty analysis in well log classification by Bayesian long short-term memory networks // Journal of Petroleum Science and Engineering. 2021. V. 205. P. 108816. DOI: <https://doi.org/10.1016/j.petrol.2021.108816>.

14. A Comparison of machine learning algorithms in predicting lithofacies: Case studies from Norway and Kazakhstan / T. Merembayev, D. Kurmangaliyev, B. Bekbauov, Y. Amanbek // Energies. 2021. V. 14. №. 7. P. 1896. DOI: <https://doi.org/10.3390/en14071896>.

15. Imamverdiyev Y., Sukhostat L. Lithological facies classification using deep convolutional neural network // Journal of Petroleum Science and Engineering. 2019. V. 174. P. 216-228. DOI: <https://doi.org/10.1016/j.petrol.2018.11.023>.

16. Dunham M. W., Malcolm A., Kim Welford J. Improved well-log classification using semisupervised label propagation and self-training, with comparisons to popular supervised algorithms // Geophysics. 2020. V. 85. №. 1. P. O1-O15. DOI: <https://doi.org/10.1190/geo2019-0238.1>.

17. Hall B. Facies classification using machine learning // The Leading Edge. 2016. V. 35. №. 10. P. 906-909. DOI: <https://doi.org/10.1190/tle35100906.1>.

18. Lithology prediction using well logs: A granular computing approach / T. M. Hossain [et al.] // Int. J. Innov. Comput. Inf. Control. 2021. V. 17. №. 1. P. 225-244. DOI: <https://doi.org/10.24507/ijic.17.01.225>.

19. Martin T., Meyer R., Jobe Z. Centimeter-scale lithology and facies prediction in cored wells using machine learning // Frontiers in Earth Science. 2021. V. 9. P. 659611. DOI: <https://doi.org/10.3389/feart.2021.659611>.

20. Analysis of ensemble methods applied to lithology classification from well logs / V. R. Leite, P. M. C. Silva, M. Gattass, A. C. Silva // 13th International Congress of the Brazilian Geophysical Society & EXPOGEF (Rio de Janeiro, Brazil, 26–29 August 2013). Society of Exploration Geophysicists and Brazilian Geophysical Society, 2013. P. 949-952. DOI: <https://doi.org/10.1190/sbgf2013-196>.

21. Ippolito M., Ferguson J., Jenson F. Improving facies prediction by combining supervised and unsupervised learning methods // Journal of Petroleum Science and Engineering. 2021. V. 200. P. 108300. DOI: <https://doi.org/10.1016/j.petrol.2020.108300>.

22. Interpreting the subsurface lithofacies at high lithological resolution by integrating information from well-log data and rock-core digital images / J. Jeong [et al.] // Journal of Geophysical Research: Solid Earth. 2020. V. 125. №. 2. P. e2019JB018204. DOI: <https://doi.org/10.1029/2019JB018204>.

23. Petrophysical characterisation of the Neoproterozoic and Cambrian successions in the Officer Basin / L. Wang [et al.] // The APPEA Journal. 2022. V. 62. №. 1. P. 381-399. DOI: <https://doi.org/10.1071/AJ21076>.

24. Hallam A., Mukherjee D., Chassagne R. Multivariate imputation via chained equations for elastic well log imputation and prediction //Applied Computing and Geosciences. 2022. V. 14. P. 100083. DOI: <https://doi.org/10.1016/j.acags.2022.100083>.
25. Lopes R. L., Jorge A. M. Assessment of predictive learning methods for the completion of gaps in well log data //Journal of Petroleum Science and Engineering. 2018. V. 162. P. 873-886. DOI: <https://doi.org/10.1016/j.petrol.2017.11.019>.
26. Synthetic sonic log generation with machine learning: A contest summary from five methods / Y. Yu [et al.] //Petrophysics. 2021. V. 62. №. 04. P. 393-406. DOI: <https://doi.org/10.30632/PJV62N4-2021a4>.
27. Sonic Waves Travel-time Prediction: When Machine Learning Meets Geophysics / W. K. Wong, Y. Nuwara, F. H. Juwono, F. Motalebi, //2022 International Conference on Green Energy, Computing and Sustainable Technology (GECOST) (Miri Sarawak, Malaysia, 26-28 October 2022). IEEE, 2022. P. 159-163. DOI: <https://doi.org/10.1109/GECOST55694.2022.10010361>.
28. Application of conditional generative model for sonic log estimation considering measurement uncertainty / J. Jeong [et al.] //Journal of Petroleum Science and Engineering. 2021. V. 196. P. 108028. DOI: <https://doi.org/10.1016/j.petrol.2020.108028>.
29. CoreScore: a machine learning approach to assess legacy core condition / M. Fellgett [et al.] //Geological Society, London, Special Publications. 2024. V. 527. №. 1. P. SP527-2021-200. DOI: <https://doi.org/10.1144/SP527-2021-200>.
30. Framework for automatic globally optimal well log correlation / O. Datskiv [et al.] //Neural Information Processing Systems (NeurIPS) Workshop on AI for Earth Sciences. 2020. P. 1-5.