Assisted Cement Log Interpretation using Machine Learning

Erlend Magnus Viggen, Norwegian University of Science and Technology; Bjørn-Jostein Singstad, Simula Research Laboratory; Eirik Time, Siddharth Mishra and Eirik Berg, Equinor;
Keywords: Machine learning; Cement log interpretation; Well integrity

Summary

The Assisted Cement Log Interpretation project has used machine learning (ML) to create a tool that interprets cement logs by predicting a predefined set of annular condition codes used in the cement log interpretation process.

The development of a cement log interpretation tool speeds up the log interpretation process and enables expert knowledge to be efficiently shared when training new professionals. By using high-quality and consistent training data sets, the project has trained a model that will support unbiased and consistent interpretations over time.

The tool consists of a training and a prediction tool integrated with the cased hole logging interpretation software. By containerizing the code using an "API First" design principle (API: Application Programming Interface), the applicability of this add-on tool is broad. The ML model is trained using selected and engineered features from cement logs, and the tool predicts an annular condition code according to the cement classification system for each depth segment in the log. The interpreters can easily fetch a complete cement log interpretation prediction for the log and use that as a template for their final interpretation. The ML model can easily be retrained with new data sets to improve accuracy even further.

To improve cement log interpretation consistency in the industry, the code will be made available as open source.

Introduction

Cementing is a critical procedure in the well construction process. The annular cement sheath provides multiple benefits (Montgomery 2006, Piot and Cuvillier 2006): It can be used to provide an annular barrier, which is a hydraulic seal to prevent undesired migration of fluids between formations and to surface. Furthermore, it anchors and mechanically supports the casing and wellhead.

After cementing, logging tools are used to check the quality of the cement. The tools predominantly used for this purpose are acoustic (Allouche et al. 2006). The simplest of these are sonic tools, where an omnidirectional transmitter excites a wave that propagates along the casing and where two receivers further down the tool then record wavefronts from the casing wave. The recordings provide a cement bond log (CBL) curve as a function of well depth, which provides information on the bonding between the casing and annular solids (Anderson and Walker 1961, Grosmangin et al. 1961, Pardue et al. 1963, Allouche et al. 2006). Low CBL values indicate good azimuthal coverage of bonded solids, while high CBL values indicate little to no coverage. (More advanced sonic tools exist, but were not used for the logs underlying this work.)

Having omnidirectional transmitters, CBL tools cannot provide any information on the azimuthal distribution of annular solids outside the casing. For that reason, ultrasonic pulse-echo tools were introduced. The principle of the tools underlying this work has been explained in detail by Hayman et al.

Author Accepted Manuscript version of the paper by Viggen et al. in *SPE Drilling & Completion*, https://doi.org/10.2118/209529-PA

(1991) and Allouche et al. (2006). In short, a rotating transducer fires ultrasonic pulses at points on the casing. At each point, processing of the recorded echo provides an estimate of the annular material's acoustic impedance, which is the product of mass density and sound speed. Hence, the resulting depth-by-azimuth impedance images allow estimating the spatial distribution of very-low-impedance gases, low-impedance liquids, and high-impedance solids. Ultrasonic pitch-catch tools were also introduced more recently by van Kuijk et al. (2005). These rotating tools comprise an angled transmitter that excites flexural waves on the casing and two angled receivers that estimate these waves' attenuation. Such estimates can be combined with pulse-echo-based estimates of acoustic impedance to help differentiate between higher-impedance liquids and lower-impedance solids (van Kuijk et al. 2005, Kalyanraman et al. 2021).

The recorded cement log primarily consists of data channels, each containing one type of tool measurement indexed against depth. Some channels, such as CBL, are represented as curves, while others, such as acoustic impedance, are represented as images. This set of channels must be interpreted as a whole to classify contiguous depth intervals in the log according to their isolating potential. The cement log interpretation process requires specialized skills, and with increased workload due to higher demand, companies may experience scarcity of such competence. In addition, cement log interpretation is difficult and subjective, and it can be challenging to get consistent interpretations over time and across interpreters (Viggen et al. 2020).

To assist professional well log interpreters in performing unbiased, consistent, and high-quality cement log interpretations over time, the project has developed an assisted cement log interpretation tool using machine learning. The trained machine learning model assists the interpreters by predicting cement log interpretations with confidence values. The predicted interpretations will provide decision support for the interpreters by acting as a starting point for their interpretation, and as such this tool can:

- Make the interpreters' jobs easier
- Help to make interpretations more consistent across multiple interpreters
- Efficiently share the expert knowledge learned by the machine learning model from a consistent, high-quality labeled training data set

It should be noted that the tool should not be used in a fully automated workflow. Cement log interpretation is a safety-critical task, and as such a professional interpreter needs to make the final qualitative interpretation decision for each cement job.

For such a tool to be useful, it must be sufficiently accurate, especially in consistently identifying zones with a high probability for isolation, as required by the NORSOK D-010 (2021) standard. False negatives (omitting intervals that could provide isolation) may be somewhat problematic, as they reduce the length of an isolating interval. False positive predictions, however, are more problematic since annular barrier quality and length will be overestimated.

The research field of producing automatic interpretations of cement logs by means of machine learning is gaining increasing interest from researchers. Since 2020, several papers on this specific topic have been published: Reolon et al. (2020) defined classes of annular conditions via clustering CBL and solid/liquid/gas fractions derived from acoustic impedance data and used Bayesian statistics to produce a probability distribution for each condition throughout each log. However, they could only report qualitative results due to the lack of a reference interpretation. Viggen et al. (2020) treated performing supervised learning (Kotsiantis et al. 2007) on expert-interpreted logs similarly to an image classification problem, using convolutional neural networks (CNNs) (Krizhevsky et al. 2012, Chollet 2018), finding human-comparable quantitative performance. Viggen et al. (2021) further improved this performance by reducing the negative combined effect of interpreter subjectivity and data complexity identified in their previous work by means of feature engineering (i.e., designing predictive features based on the raw log data) and using ML algorithms that are less susceptible to overfitting (Webb 2011). Voleti et al. (2020) also reported good results with a comparable tool trained and tested on a small number of wells, estimating that it could save their company 75% of their current interpretation effort.

Furthermore, other related work has been reported: Kalyanraman et al. (2021) discussed how errors in the processing of ultrasonic pulse-echo and pitch-catch measurements can be corrected (either manually or via ML) to facilitate a more robust and detailed characterization of the annular materials behind the casing prior to interpretation. Imrie (2021) developed an algorithm to distinguish between good cement coverage, partially good bonding quality cement coverage, and free pipe using CNNs on acoustic impedance maps. Furthermore, he also developed a method to detect and classify fluid channels, in cases with partial coverage, into four predefined classes.

The work reported here merged an operator's internal R&D project with published academic methods (Viggen et al. 2021) to develop a machine learning tool that will be used by interpreters in real-world cases. While the tool is currently limited to usage with the logging data of one service company, it is written with extensibility in mind. The tool is written in Python and is currently in a beta state where it is being tested in practice by cased hole logging domain experts. When the tool is completed, it will be released under an open-source license to benefit the rest of the industry.

Methods

The general approach has been to use supervised machine learning, providing a machine learning algorithm examples of well log data and corresponding interpretations, so that it can learn generalizable relationships between these. The machine learning pipeline operates on the numerical log data rather than plotted log curves and images. It can therefore not be biased by plot style choices such as log images' color maps.

Dataset

A supervised machine learning approach requires raw data and interpretations as the basis for training and testing. The dataset underlying this work consists of 70 interpreted cement logs on the standard DLIS well log format defined in API RP66 V1 (1991), mostly gathered between 2017 and 2021. These DLIS files are read using dlisio, an open-source Python library (Equinor 2019). Every log contains a CBL channel and an acoustic impedance channel, and most of the logs contain channels from pitch-catch tools. All log files are preprocessed to azimuthally rotate depth-by-azimuth image data so that the middle of the image corresponds to the bottom of the casing, and to reduce the file size by discarding data channels that are not relevant for interpretation.

The logs are all interpreted according to the same interpretation schema: The interpretation process splits the logged well section by depth into consecutive zones according to the condition of the materials in the annulus. Each zone is assigned an *annular condition code* from a set of codes defined based on experience gathered from impedance distributions in sonic and ultrasonic data. The full list of the 30 annular condition codes used in this project can be found in Table 1. In particular, note that the interpretation schema includes MAWCem/MADCem codes representing various types of microannuli — gaps between casing and cement ranging from a few µm to hundreds of µm (Jutten and Hayman 1993, Kalyanraman et al. 2017, Issabekov et al. 2017) — and the FORMCEM codes representing combinations of cement and formation, typically where a creeping formation has closed in on the casing through the gaps in cement coverage (see e.g. Kalyanraman et al. 2021 for an example).

Table 1—The 30 annular condition codes used for interpretation

Code	Isolating potential	Description
CEM 1A	High	Well Bonded, homogeneous, cement around the entire annulus
CEM 1B	High	Well Bonded, heterogeneous, cement around the entire annulus
CEM 1C	High	Well Bonded, heterogeneous, cement around the entire annulus with non-connected small liquid pockets / short liquid filled channels
CHNCem 2A	Low	Continuous Liquid Filled Channeling — <20% Fluid Channel Width
CHNCem 2B	Low	Continuous Liquid Filled Channeling — 20–40% Fluid Channel Width
CHNCem 2C	Low	Continuous Liquid Filled Channeling — >40% Fluid Channel Width
CONCem 3A	High	Well bonded Cement High Side, Slightly Contaminated Cement Low side
CONCem 3B	Medium	Well bonded Cement High Side, Heavily Contaminated Cement Low side

CONCem 3C	Medium	Homogeneous Contaminated (or Unset) Cement — Lower than the expected impedance
PATCem 4A	Medium	Patchy Cement Bond — Medium Isolating potential
PATCem 4B	Low	Patchy Cement Bond — Low Isolating potential
MAWCem 4C	High	Wet Microannulus — High Isolating Potential
MAWCem 4D	Medium	Wet Microannulus — Medium Isolating Potential
MAWCem 4E	Low	Wet Microannulus — Low Isolating Potential
MADCem 4F	Low	Dry Microannulus
GCCem 5	Low	Gas Cut Cement
LLCem 6	Low	Cement in Liner Lap — Eccentralized Liner
FORM 7A	High	Formation Bond Good / Barrier Quality
FORM 7B	Medium	Formation Bond Medium / Not Barrier Quality
FORM 7C	Low	Formation Bond Low / Not Barrier Quality
FORMCEM7C	Low	Combined Cement and Formation Bond — Poor Bond
FORMCEM7D	Medium	Combined Cement and Formation Bond — Medium Bond
FORMCEM7E	High	Combined Cement and Formation Bond — Good Bond — PRESSURE TEST REQUIRED
MUDS 8A	Not applicable	Settled mud solids — High Density / Well Bonded
MUDS 8B	Not applicable	Settled mud solids — Medium Density / Medium Bond
MUDS 8C	Not applicable	Settled mud solids — Low Density / Patchy Bonded
FPL 9A	Not applicable	Liquid Filled Free Pipe
FPG 9B	Not applicable	Gas (or light oil) filled free pipe
OTHER		If the annulus status cannot be classified by one of the above categories
PDQ		Data Quality Issues ——such that data is not interpretable

Furthermore, for each log, the dataset contains essential external information that is not provided in the original log file, such as:

- The theoretical top of cement calculated from the borehole diameter, the casing diameter, and the pumped cement slurry volume
- Whether losses occurred during cement displacement (Daccord et al. 2006)
- The time elapsed between cementing and logging
- The depth of the outer casing shoe

Previous work has shown that keeping interpretations consistent is essential for a good ML result (Viggen et al. 2020, Viggen et al. 2021). Therefore, the cement logs in the dataset have been hand-picked by a team of cement log interpretation experts, who have performed an additional quality control (QC) step to further improve the internal consistency of the cement logs in the dataset. We describe the construction of the dataset in more detail in the Data Selection section below. Even so, maintaining perfect consistency is not possible. There are various uncertainties related to the condition of the downhole environment that relate to the cement job and hence the interpretation of the cement log. Some aspects of the interpretation process are also subjective, for example, what level of detail is appropriate. Where one interpreter defines one long zone, another interpreter could prefer to define several smaller zones.

Table 2—Distribution of the annular condition codes defined in Table 1, showing the number of 1 m segments and the number of logs with each annular condition code, across the training and test sets

Annular condition code	Training set		Test set	
	No. of segments	No. of logs	No. of segments	No. of logs
FPL 9A	18 101	34	6026	15
FORM 7A	5846	18	2006	7
CEM 1A	4957	21	1931	10
Other	4720	28	1215	9
FORM 7B	3436	14	1706	10
CEM 1C	2476	27	669	12
CONCem 3A	2442	16	735	7
CEM 1B	2343	22	889	10
MUDS 8C	2288	11	693	3
FORM 7C	2165	11	1853	8
CONCem 3B	1876	17	723	8
CHNCem 2B	1780	15	644	9
PATCem 4B	1774	15	505	9
MUDS 8B	1227	4	258	3
PATCem 4A	978	15	274	7
FPG 9B	899	6	4	1
MUDS 8A	880	5	256	2
FORMCEM7D	852	2	160	4
CHNCem 2C	749	6	295	2
LLCem 6	705	4	30	1
PDQ	635	7	560	2
FORMCEM7C	545	2	363	1

CHNCem 2A	538	9	244	4
CONCem 3C	467	7	207	1
MADCem 4F	420	2	0	0
MAWCem 4C	296	5	236	1
MAWCem 4D	240	3	161	2
FORMCEM7E	205	2	35	2
GCCem 5	149	2	236	1
MAWCem 4E	118	2	0	0
Total	**64 107**	—	**22 914**	—

Following common practice in ML, the dataset is split into two parts, one for training the algorithm, and one for testing it on unseen data. The split was performed on a log-by-log basis in order to avoid similar data from the same log being present in both parts, as this would lead to unrepresentative improved test performance. Like Viggen et al. (2020) and Viggen et al. (2021), we use an optimization algorithm based on simulated annealing (Press et al. 2007) to ensure that the two parts have a similar distribution of annular condition codes. The training set consists of 50 logs, and the test set consists of 20 logs. Table 2 shows how the different annular condition codes are distributed throughout the training and test sets.

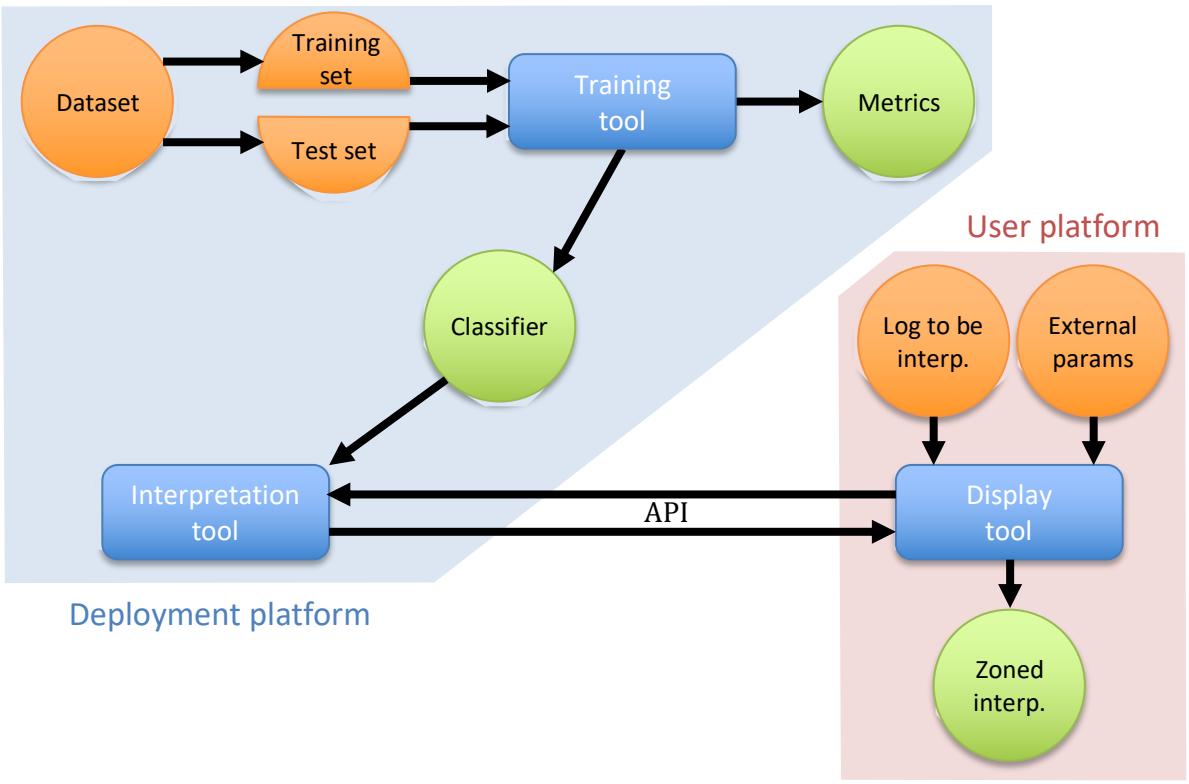

Figure 1—Simplified overview of the structure of the automatic cement log interpretation system

Data Selection

The cement log dataset underlying this work was built in a step-by-step fashion. Initially, it consisted of a small number of logs with clear-cut interpretations to ensure that the ML had good examples of the most common codes. As the project evolved, logs of a more typical quality and level of uncertainty were added. To guide the selection of new logs, the distribution of the annular condition codes (as seen in Table 2 for the dataset reported here) was continuously monitored to identify annular condition codes with insufficient representation in the dataset. Furthermore, logs that were reviewed and interpreted as part of the day-to-day activities of the company's cased hole logging experts were also added to the dataset after an additional QC step. However, logs that were identified as very difficult to interpret due to high levels of the aforementioned uncertainties were not included in the dataset; such logs would be poor training examples or poor references for testing a trained system.

In order to reduce the effect of interobserver variation (Popović and Thomas 2017) of the cement log interpretations, previously interpreted logs were subject to an additional QC step. In this process, a small team of two interpreters fine-tuned previous interpretations to further improve the consistency in these interpretations' level of detail and choice between bordering annular condition codes.

The QC step also identified logs where the acoustic impedance image channels showed spurious stripes of low impedance caused by mud deposits on the bottom side of the casing in deviated wells (Thierry et al. 2016) and ensured that more suitable acoustic impedance channels, produced by reprocessing the underlying ultrasonic pulse-echo data with a processing algorithm that accounts for this effect (Klieber and Lemarenko 2016), were available to the ML before including the log in the dataset.

System Structure

The automatic cement log interpretation system can be viewed as consisting of three interlinked components, as shown in Figure 1: A training tool, an interpretation tool, and a display-tool.

Training Tool

Using the training set of interpreted cement log data, the training tool trains an ML classifier to be able to interpret input cement log data. If the training is successful, the classifier can provide reasonable interpretations of unseen cement log data. To test that the training process is successful, the training tool can calculate various performance metrics. These metrics can be calculated on the test set, as well as on the training set through k-fold cross-validation (Refaeilzadeh et al. 2011, Viggen et al. 2020). Figure 2 shows an overview of the training tool, including details that will be explained in later sections.

In the cross-validation process, the logs in the training set are distributed into k different folds. As before, the simulated annealing-based algorithm is employed to distribute the logs while optimizing the balance of annular condition codes between the folds. Each fold is, in turn, used for testing after the remaining k–1 folds have been used for training a classifier. In the end, the test results for all the folds are combined to calculate the desired performance metrics.

When the tool is in regular use, the trained classifier will regularly be retrained with additional logs and interpretations added to the training set. It is therefore expected that the interpretation tool's performance will gradually improve over time.

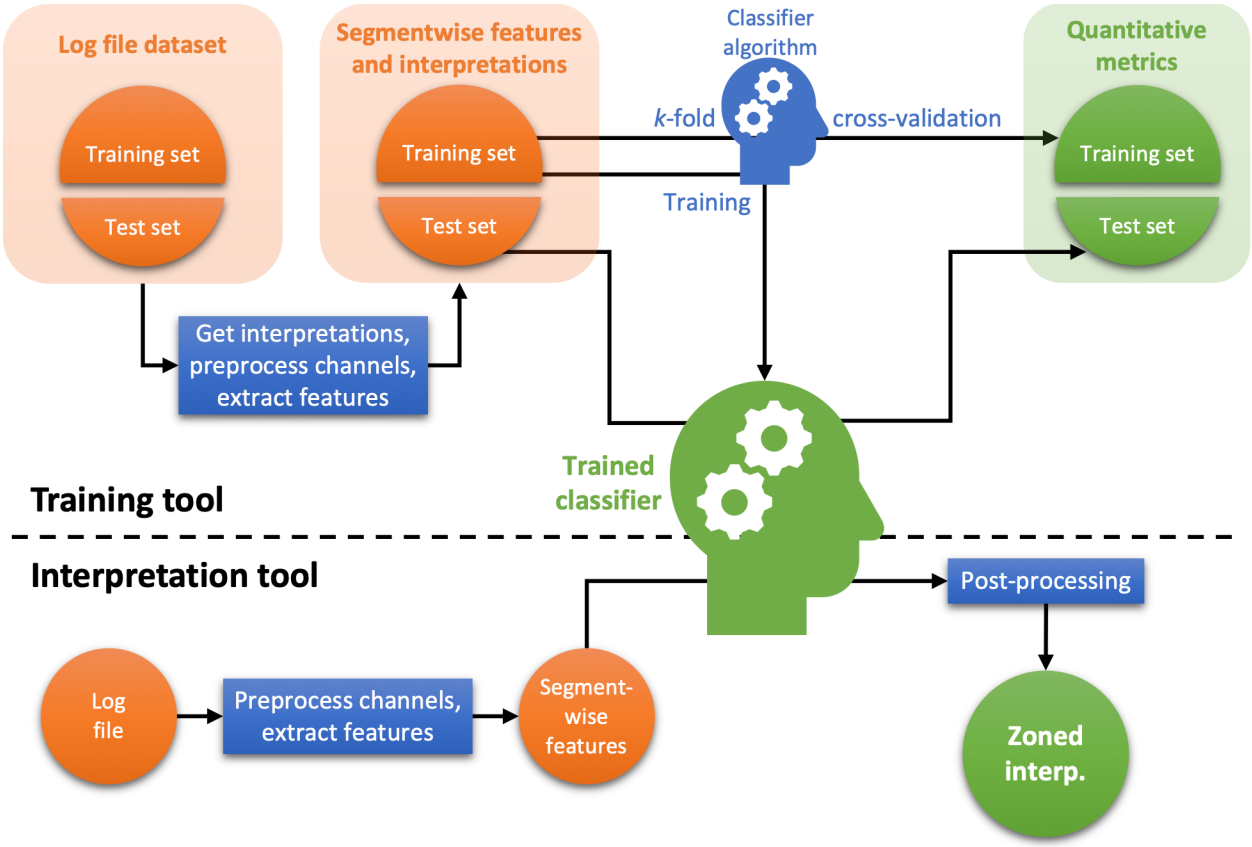

Figure 2 - Detailed overview of the training and interpretation tools

Interpretation Tool

As Figure 2 shows, when the interpretation tool is provided with a classifier from the training tool and an appropriately preprocessed cement log file, it produces a log interpretation on the same zonation form used by the expert interpreters. Furthermore, it provides a confidence curve specifying at each depth how confident the classifier is about its conclusion, based on the classifier's class probability distribution. Even though such a confidence curve may not be very predictive of where classifiers make errors (Viggen et al. 2020), the curve helps interpreters see where the classifier is close to making another interpretation. To further help the interpreters, the tool also provides the *second* most likely annular condition code for each zone, according to the classifier.

To help avoid getting many very short zones as seen in previous publications on automatic cement log interpretation tools (Reolon et al. 2020, Viggen et al. 2020, Voleti et al. 2020, and Viggen et al. 2021), the tool includes a post-processing component. The user can specify the minimum length of a zone (e.g., 3 m), and the tool will merge and/or replace shorter zones based on the classifier's class probability distribution. The post-processing also assists the machine learning by removing impossible annular conditions from consideration, such as the presence of formation in dual-casing geometries and free pipe sections with gas or light oil zones below zones of heavier liquids.

The interpretation tool exposes an application programming interface (API), to ensure that it can be treated as a separate component and integrated with a variety of programs as necessary. The API can be provided with a log file and outputs a zoned interpretation, an alternate interpretation, and a confidence curve.

Display-Tool

The most common use case is to use machine learning interpretation systems directly from the cased hole logging interpretation software used by the professionals. The predicted interpretation is plotted alongside the log data, and the professional will then correct the automatic interpretation as necessary. The display-tool in this work is based on a pre-existing commercial log-interpretation software that also provides Python scripting capabilities.

When the interpreter opens a new cement log and wants an automatic interpretation, a Python script in the display-tool gathers the necessary log data channels and parameters, writes them to a temporary log file, and calls the interpretation tool's API to point it to the temporary file and the classifier to be used. The API returns a zoned interpretation, in addition to the confidence curve and alternate interpretation, which the interpreter can then display and edit. This process is described in more detail in the Using the Interpretation Tool section below.

Machine Learning Task

To build an ML-based automatic interpretation system, the problem of interpreting cement logs must be posed in a fashion that is appropriate for machine learning. The project has utilized the same approach as Viggen et al. (2020) and Viggen et al. (2021). The well is split into depth segments of 1 meter length, and the classifier's task is to provide an annular condition code for each of these segments based on features extracted from larger surrounding intervals of well log data.

The classifier is not tasked to recognize every single annular condition code in the interpretation schema; two annular condition codes have been excluded, namely "Other" and "PDQ" from training. "Other" is a catch-all code for different conditions that cannot be generalized and are not very important (specified in a comment field in the final interpretation), and its diversity makes it difficult for a classifier to learn. The "PDQ" code indicates poor quality of the log data, which is usually easy enough for human interpreters to spot but can be more difficult for a classifier to identify.

Data Preprocessing

It is well-known, and evident from any acoustic well log plot, that acoustic well log data contains systematic outliers around casing collars and casing centralizers. This is because the acoustic logging

techniques implicitly assume the casing to be continuous and uninterrupted. For example, CBL curves contain spikes near casing collars, and acoustic impedance estimates from ultrasonic pulse-echo measurements contain spurious high values at casing collars and spurious low values at casing centralizers. While human interpreters quickly learn to look past these outliers, ML algorithms generally take their inputs at face value. Therefore, the effect of the outliers should not be present in the ML input.

For CBL data, a median filter was sufficient to remove the effect of the casing collar spikes, as we explain in the Feature Extraction section below. To extract precise features from the image channels produced by the ultrasonic tools, however, the outliers should first be removed from the images.

At a casing collar or centralizer, the resulting outliers are visible in the image data for most, if not all, of the azimuths over a small range of depths. To identify these depths, the mean value of the image was first taken over all azimuths, providing a curve against depth where strong spikes at collars or centralizers stick out from an otherwise relatively smooth curve. Applying a median filter with a kernel size of 1.5 m to this curve provides a slightly smoother version of the same curve, without the spikes. The depths containing the spikes can then be identified by thresholding the difference between the original curve and the median filtered curve. All values at these depths in the image channel are replaced with special values called Not a Number (NaN) to indicate the absence of usable values. These NaN values are ignored when calculating the ML input as described in the Feature Extraction section.

Furthermore, some scattered outliers are caused by issues with the recording or processing of ultrasonic pulses. A separate image channel flags pulses with such issues, and values derived from the flagged pulses are also replaced with NaN.

Feature Extraction

Another essential part of posing the ML problem is to determine which form the input cement log data should take when provided to the classifier. One approach, taken by Viggen et al. (2020) and the operator's internal R&D project, is to treat the classification similarly to a computer vision problem. Here, CNNs (Krizhevsky et al. 2012, Chollet 2018) automatically determine the connections between the log data and outputs. This work selected, however, the feature engineering approach that Viggen et al. (2021) showed to be superior to the former approach on the same dataset.

In the feature engineering part, the project designed a set of features that could be calculated based on the cement log data. In this manner, the domain knowledge can be used to design features that are relevant for cement log interpretation. Ideally, each feature should be directly predictive of one or more classes. For each 1 m segment to be interpreted, all features are calculated, typically from a “context interval” of surrounding well log data. The length of the context interval depends on the type of data the feature is based on. Where systematic outliers are not removed as described in the Data Preprocessing section, the context interval must be long enough to avoid the impact of such outliers, while remaining short enough that short-scale cement log structures such as small fluid patches can be detected. In addition to these shorter context intervals, each feature based on log data is also recalculated for a long interval of 20 m length, in order to provide longer-range context as suggested by Viggen et al. (2021).

The choice of features is largely similar to that of Viggen et al. (2021), where they are more rigorously defined than in the following summary. In machine learning, it can be critical to ensure that all features have a similar range of values. Hence, every feature is normalized to a value interval $[0,1]$ based on its distribution in the dataset.

Acoustic Impedance Features

Many of the features are based on the aforementioned depth-by-azimuth acoustic impedance images. Because the preprocessing removes the vast majority of the systematic outliers, a context interval as short as 2 m could be used. The features calculated from each context interval of the preprocessed impedance image data are:

- The median impedance in the interval.

- The impedance heterogeneity, which is calculated as the absolute difference between the median impedance and a particular impedance percentile. In this work, the 10th and 90th percentiles are used. This feature helps quantify the spread of impedance values.
- The difference between the median impedances in the top $(360° - \theta)$ and the bottom θ azimuths of the casing. In this work, two such features are calculated, one for $\theta = 180°$ to help identify wide structures at the bottom of the casing, and one for $\theta = 40°$ to help identify narrower structures such as thin liquid channels.
- The liquid fraction, defined as the proportion of impedances in the interval between predefined gas-liquid and liquid-solid thresholds, typically 0.3 MRayl and 2.6 MRayl, respectively.
- The solid fraction, defined as the proportion of impedances above the liquid-solid threshold.

CBL Feature

Furthermore, a feature based on the cement bond log (CBL) curve is calculated, which has high values in free-pipe sections (where the impedance is low) and low values in well-cemented sections (where the impedance is high). To suppress the outlier spikes in the CBL curves around casing collars, a 4 m context interval has been used. Because the expected CBL value in free-pipe sections varies with the casing size, the CBL values are not necessarily comparable across different logs. To compensate, we normalize each log's CBL curve with the expected free-pipe CBL value for that log. While Viggen et al. (2021) found that CBL did not provide a significant additional benefit to impedance-based features, that work was based on a different interpretation schema that did not take microannuli explicitly into account like the schema used here does with the MAWCem/MADCem codes shown in Table 1. Different types of atypical relationships between the CBL and impedance values can be used to identify and differentiate wet and dry microannuli (Jutten and Hayman 1993, Kalyanraman et al. 2017, Issabekov et al. 2017), which should make the normalized CBL curve more valuable. From this curve, only one feature is calculated:

- The median normalized CBL value in the interval

External Features

The project also used a number of features based on external information:

- From the theoretical top of cement and the presence or absence of cement losses, a curve indicating the probability of a given depth segment being below the actual top of cement is calculated, based on the distributions of theoretical and actual top of cement in the dataset.
- From the outer casing shoe depth, a binary feature specifying whether or not there is another casing outside the logged casing at the given depth segment is calculated.
- From the time between cementing and logging, the tool calculates a value correlated with the probability of formation collapse having occurred at the time of logging.

Unused Features

While the project did design more features than those mentioned above, some were rejected for various reasons, including a feature evaluation based on univariate statistical analysis and recursive feature elimination based on permutation feature importance. The rejected features were based on:

- Log data from an ultrasonic pitch-catch tool (van Kuijk 2005). Unlike the ubiquitous ultrasonic pulse-echo and CBL tools, pitch-catch tools are not provided by a wide range of service companies. Hence, relying on the presence of pitch-catch data would rule out training a single classifier that can work on logs from a range of service companies. Furthermore, feature evaluation found that the pitch-catch features, based on pitch-catch attenuation and solid-liquid-gas maps, did not improve the performance of the machine learning. This may be partly because the non-linear relationship between pitch-catch attenuation and acoustic impedance (van Kuijk 2005) is difficult for machine learning to deal with.

- Gamma ray log data. Gamma curves are difficult to normalize to make them comparable between different wells. This feature did not improve performance.
- Features based on the area and length of consecutive fluid regions. The idea of these features was to help differentiate wide but short fluid patches and long but narrow fluid channels, but these features did not improve performance.

Quantitative Metrics

To objectively evaluate the prediction performance of the ML tool, we must define some performance metrics. As there is no ground truth available, the expert interpretations must be treated as a "correct" reference. The interpretations' match with the ML predictions can be quantified on a segment-by-segment basis. For the i th 1 m segment out of all N , the chosen annular condition code (Table 2) in the expert interpretation is denoted as y_i and the annular condition code in the ML prediction as $\hat{y_i}$.

The most basic metric is *unbalanced accuracy*, which is the proportion of segments where the two annotations match:

$$UA(y, \hat{y}) = \frac{1}{N} \sum_{i=0}^{N-1} 1(y_i = \hat{y_i}).$$

Here, $1(y_i = \hat{y_i})$ is an indicator function that equals 1 if its argument is true and 0 if it is false. However, unbalanced accuracy is most heavily affected by the performance on the most prevalent annular condition codes and is hardly influenced by the performance on rare annular condition codes. The *balanced accuracy* metric weights each segment as $w_i = 1/\left[\sum_j 1\left(y_j = \hat{y_j}\right)\right]$ i.e. inversely proportional to the prevalence of its annular condition code prevalence so that every annular condition code is made equally important:

$$BA(y, \hat{y}) = \frac{1}{\sum_{i=0}^{N-1} w_i} \sum_{i=0}^{N-1} 1(y_i = \hat{y_i})$$

Accuracies calculated based on individual classes may not necessarily give the right impression. Whether the annular condition code is, e.g., CEM 1B or CEM 1C often does not matter in practice, as both annular condition codes represent the same high isolation potential indicating a barrier. For a more coarse-grained representation, the annular condition codes can be pooled according to their isolating potential. Then, unbalanced and balanced accuracies can be calculated in terms of isolating potentials instead of individual annular condition codes.

Furthermore, as mentioned in the introduction, the most important job of the ML tool is to correctly separate barrier-level zones of high isolating potential from zones that are not barrier-level. To quantify how well the ML tool performs this task, the coarse-graining can be taken further, pooling all annular condition codes of high isolation potential into a "positive" class and all other annular condition codes into a "negative" class. With two classes, this becomes a binary classification problem with four possible outcomes per 1 m segment, according to whether the ML prediction assigns the same positive or negative class as the reference expert interpretation:

- True Positive (TP) is a correct positive assignment
- True Negative (TN) is a correct negative assignment
- False Positive (FP) is an incorrect positive assignment
- False Negative (FN) is an incorrect negative assignment

Standard metrics for binary classification problems can then be calculated from the total number of these outcomes:

- Precision (TP/(TP+FP)) is the proportion of correct positive ML predictions
- Sensitivity (TP/(TP+FN)) is the proportion of positive interpretations where the ML prediction is correct

- Specificity (TN/(TN+FP)) is the proportion of negative interpretations where the ML prediction is correct
- F1 score (TP/(TP+(FP+FN)/2)) gives an overall accuracy via the harmonic mean of precision and sensitivity.

Finally, to give more insight into exactly what kind of mispredictions the ML classifier tends to make, the joint distributions of reference interpretations and ML predictions can be displayed as balanced confusion matrices (Ting 2017). The matrix element at the k th row and l th column represents the proportion of segments where the ML prediction is the l th annular condition code out of all segments where the reference interpretation is the k th annular condition code. Thus, for a hypothetically perfect prediction, only the diagonal matrix elements would be populated; however, perfect agreement is not realistic even among experts (Viggen et al. 2020). Confusion matrices can be similarly built for the coarse-grained representations as well.

Choice of Machine Learning Classifier

There are many types of machine learning classifiers for supervised learning that all do a comparable job. During training, they learn the relationship between the inputs (features) and outputs (interpretations) by example. After training, they predict outputs based on inputs that they have not seen before.

There are also many software libraries providing implementations of such classifiers. While not all previous articles on ML-based cement log interpretation specify which libraries were used, Viggen et al. (2020) used the Keras library's (Chollet et al. 2015) implementations of CNNs, and Viggen et al. (2021) showed significant improvements over these earlier CNN results with a feature engineering-based approach using the scikit-learn library's (Pedregosa et al. 2011) implementations of various classifiers. Of the 8 different classifiers that the latter article compared, logistic regression (LR) classifiers performed the best, statistically significantly outperforming most of the others. For this reason, LR classifiers were also used for this work as the setup is very similar to that of Viggen et al. (2021). However, because the improvements to the interpretation schema and data quality in this work could have affected the relative performance of classifiers, this project also re-evaluated the gradient tree boosting (GTB) classifier. According to Chollet (2018), GTB is often the best-performing classifier for "non-perceptual" problems that do not deal with image or sound data, and Viggen et al. (2021) also found GTB to be among the top-performing classifiers tested. However, an initial re-evaluation on the current dataset found that GTB is still outperformed by LR by 7% and 14.6% measured in terms of unbalanced and balanced accuracy in classifying all 28 classes and 3.6% and 2.7% measured in terms of unbalanced and balanced accuracy when determining the isolation potential. (When comparing the two, the hyperparameters of both by means of a grid search were optimized. This optimization only weakly affected the performance, indicating that the performance is mainly limited by other factors.) Hence, only LR-based results are presented in the Results section below.

Using the Interpretation Tool

A Graphical User Interface (GUI) was developed using the Python environment of the cased hole log interpretation software which allows the user to select the well, dataset, and a few critical parameters such as the theoretical top of cement, previous casing shoe, losses during cementing, etc. Once the user inputs this information using the GUI, this tool then prepares the raw log data before calling the API of the interpretation tool. Figure 3 shows the workflow.

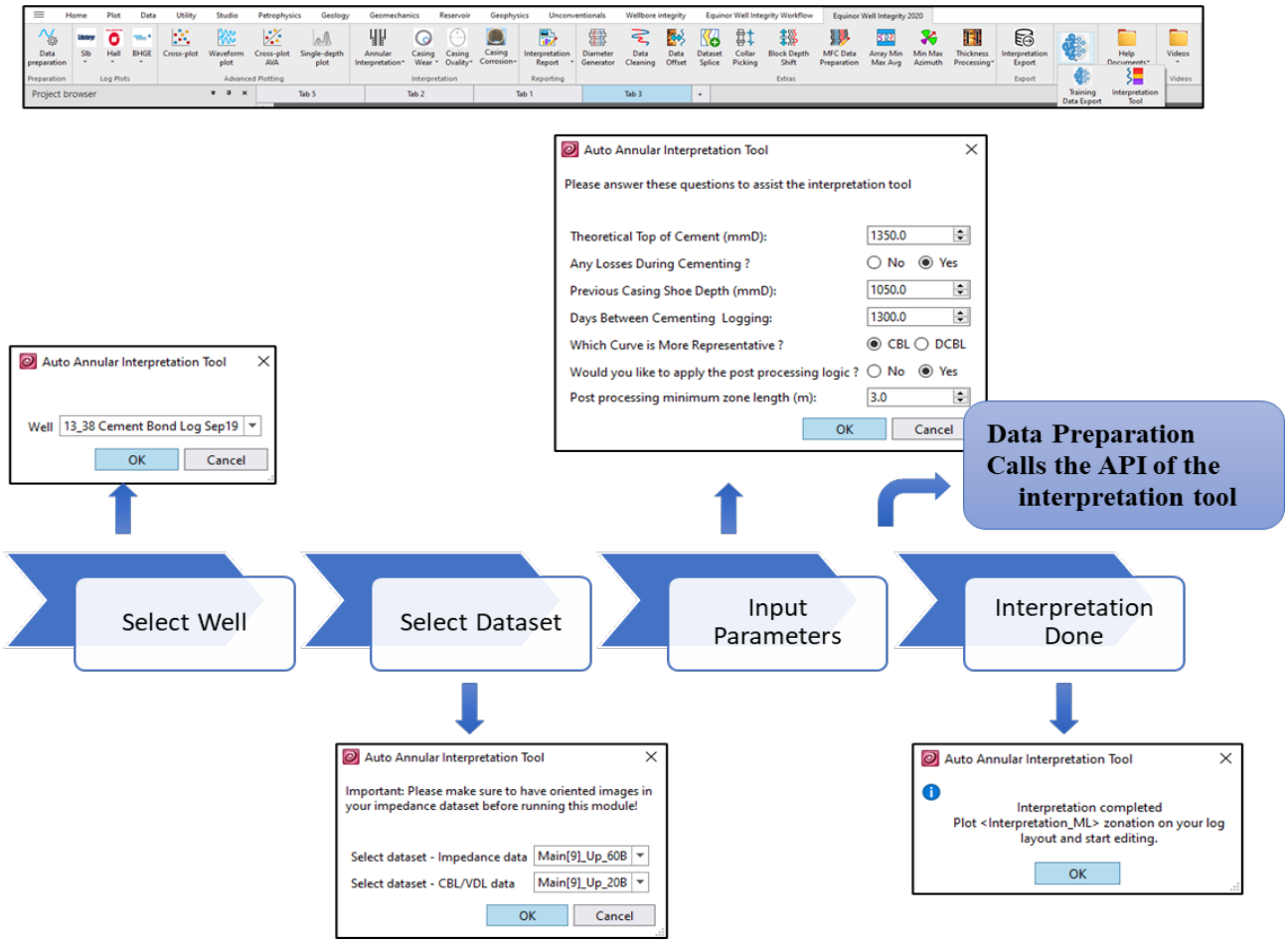

Figure 3—Workflow of calling the interpretation tool from the display-tool

Below is an example of how the output of the interpretation tool looks like (the last 3 tracks in Figure 4). Two zonation datasets are created: The first is the classifier's most confident zonation dataset with interpretation classes, and the second zonation contains the second most likely class according to the classifier. The last track in Figure 4 shows the ML interpretation's confidence curve — the confidence that its primary interpretation is correct. The last step of this entire workflow is for expert interpreters to modify these ML-based interpretation zonations by applying their domain knowledge and field expertise.

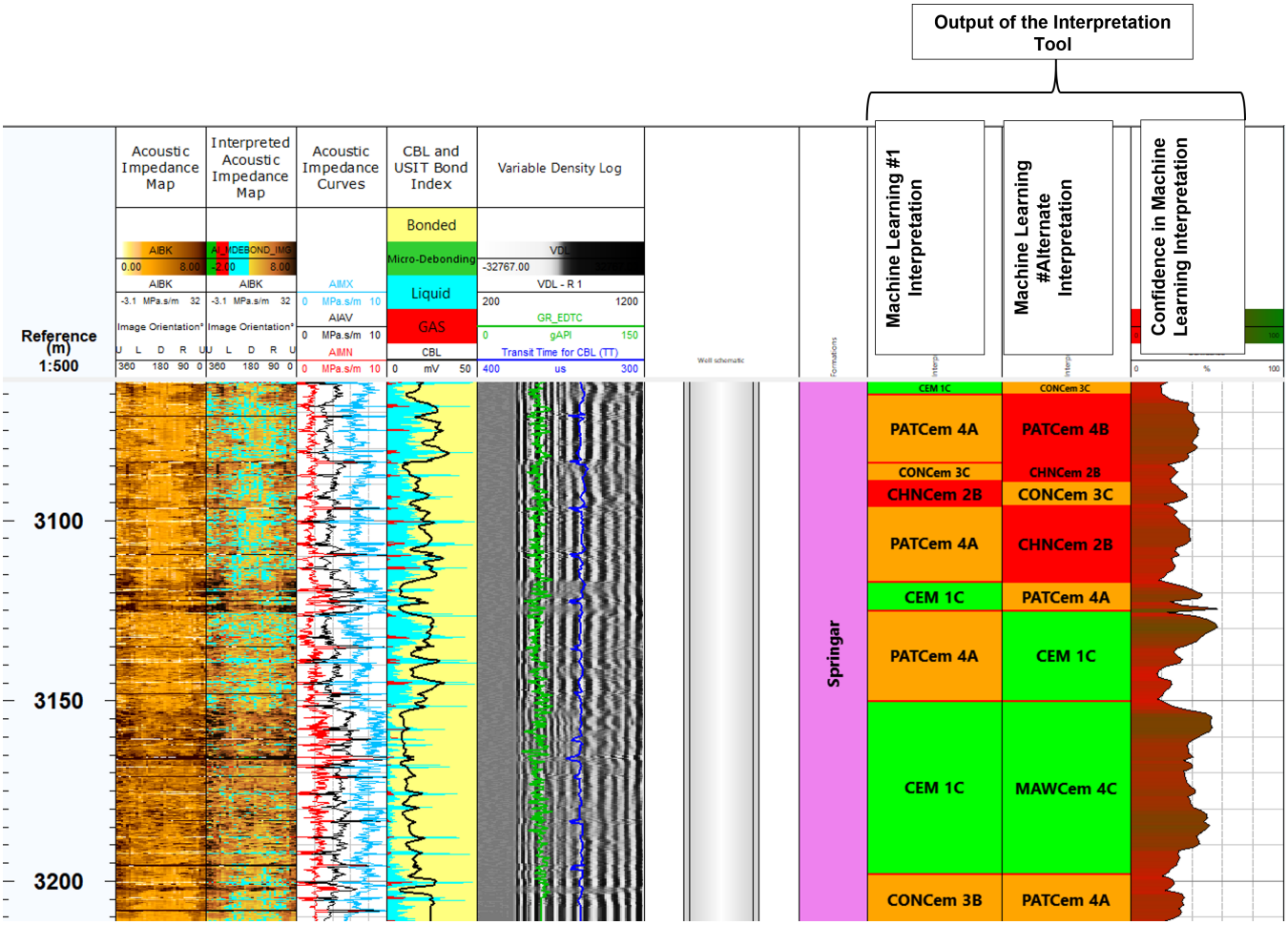

Figure 4—Example output of the interpretation tool. 1st column: Acoustic impedance behind the casing as a depth-by-azimuth image. 2nd column: The same acoustic impedance, with color thresholds indicating liquids (blue) and gas (red). 3rd column: The maximum (blue), average (black) and minimum (red) acoustic impedance value at each depth. 4th column: CBL curve (black) on top of the relative azimuthal proportion of solid (yellow), liquid (blue), and gas (red) at each depth, estimated by thresholding the acoustic impedance. 5th column: Variable density log (VDL) showing the sonic tool's recorded waveform at each depth, with gamma ray (green) and CBL transit time (blue) measurements superimposed. 6th column: Well schematic. 7th column: Formation names. 8th column: The primary prediction of the ML tool. 9th column: The alternate prediction of the ML tool. 10th column: The ML tool's confidence at each depth that the primary prediction is correct.

Results

Qualitative Results

Figure 5 and Figure 6 below show two example log datasets and compare the interpretation performed by domain experts with the ML predicted interpretation. It can be observed that ML predicts a greater number of zones as compared to the expert interpretations. However, there is a very good agreement on the isolation potential (represented by zone color) between the ML and expert interpretations for both the example datasets. Moreover, none of the zones interpreted as medium or low isolating potential by experts are predicted to have high isolating potential by the ML model. This is a very important point, because this annular bond interpretation is a safety-critical task and misinterpreting a zone to have high isolating potential can have serious consequences. The disagreement in the annular condition codes between expert and ML interpretations are discussed in more detail in the Quantitative Results and Discussion sections of this paper.

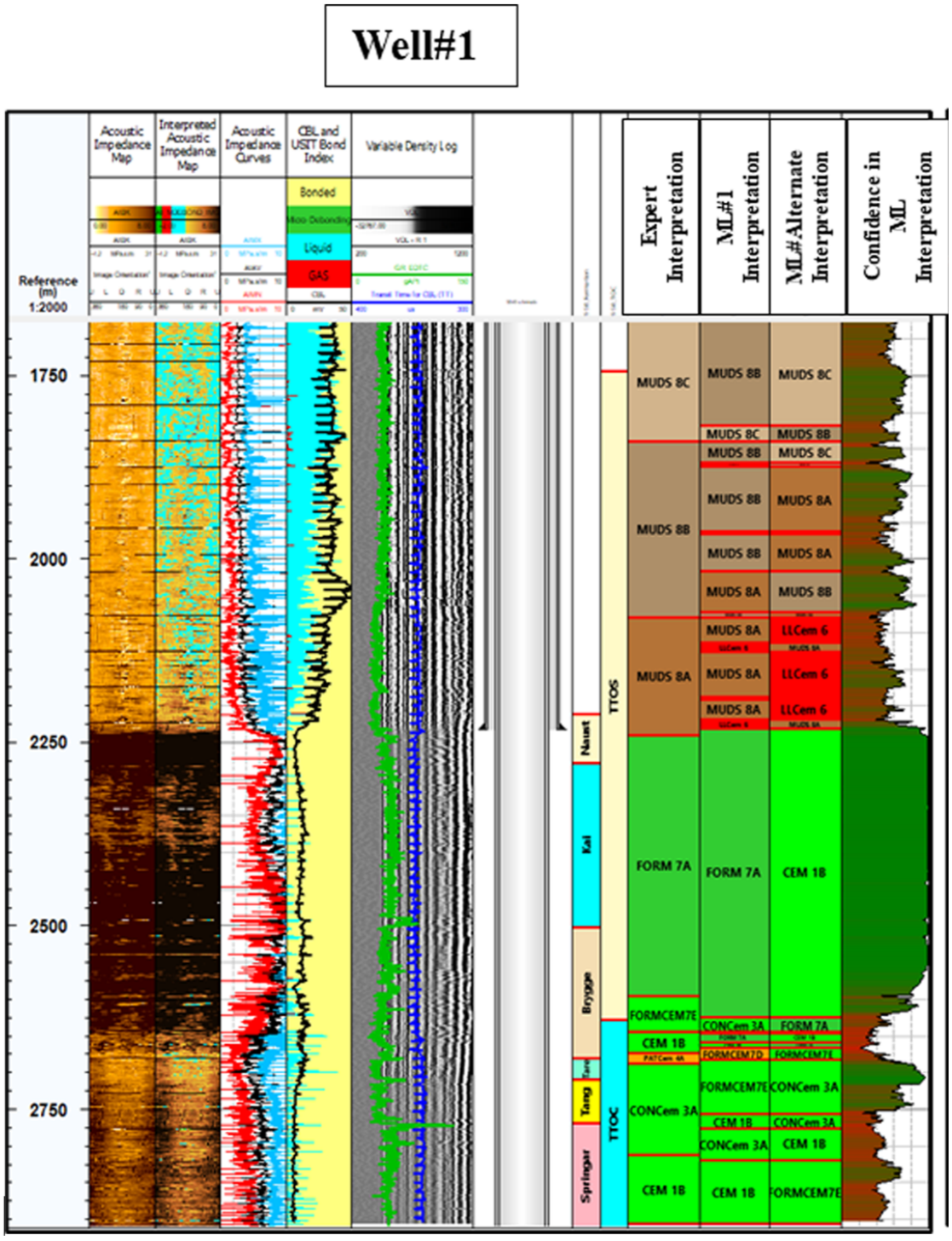

Figure 5—Comparison of expert interpretation with ML interpretation for one example log. The columns represented are plotted similarly to Figure 4 , but with the addition of a column showing the expert interpretation of the wells.

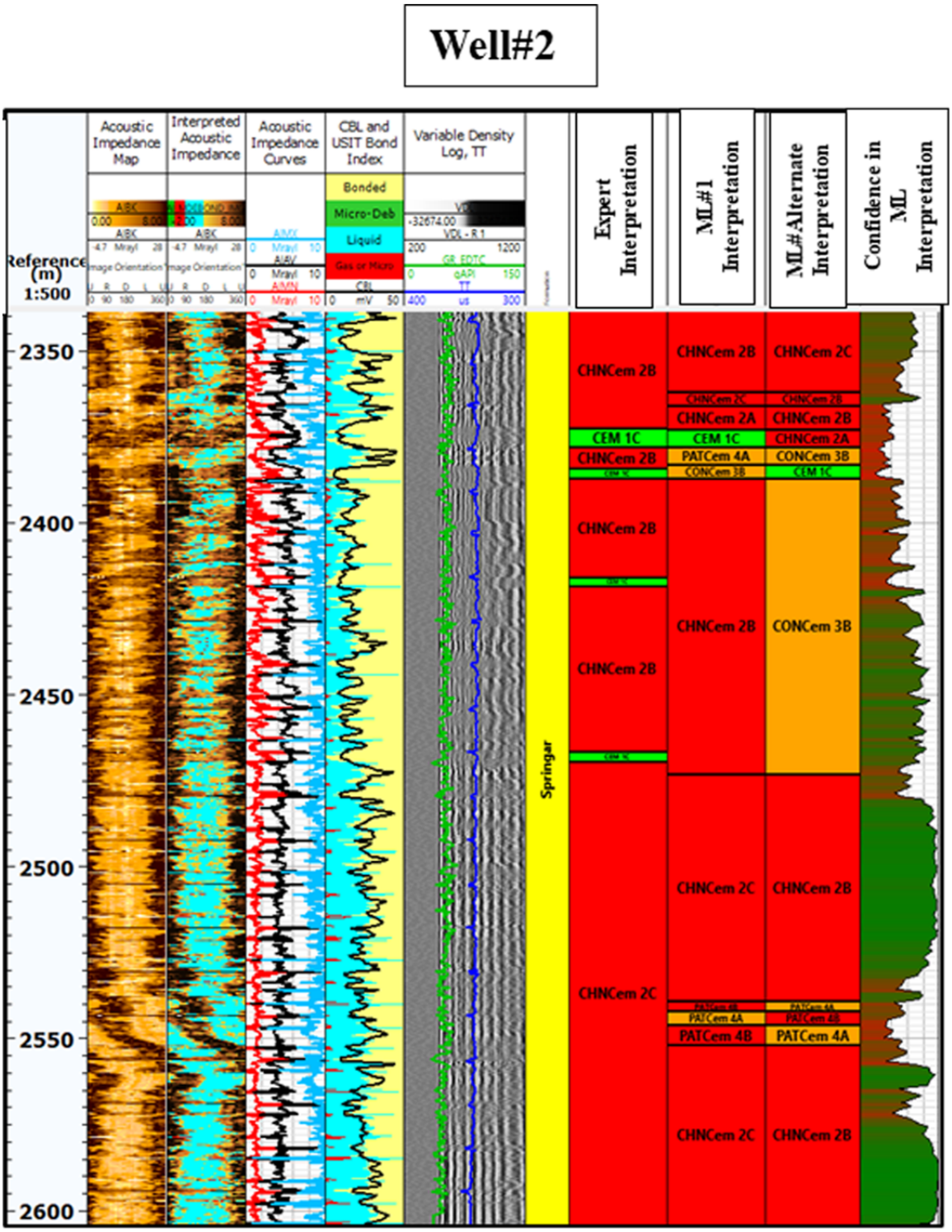

Figure 6 - Comparison of expert interpretation with ML interpretation for another example log, similar to Figure 5.

Quantitative Results

Cement log data can be used both to train the classifier and to test its performance. However, one must be careful not to test the classifier on data already used during training, only testing it on *unseen* data.

There are multiple ways to avoid this. The classic approach is to split the dataset into a training set and a test set, as described in the Dataset section above, train the classifier on the training set, and evaluate its

performance on the test set. However, if the test set is too small to reliably capture the possible variations in the dataset, the performance on the test set may not be representative of the performance on a larger dataset. Hence, one can perform *k*-fold cross-validation on the larger training set. (This work uses *k*=5.) The results of this process may be more representative as the training set is larger, though the results may also be biased if some part of the training optimization procedure has overfitted the classifier to the training set. Performing both these evaluations gives complementary results; the 5-fold cross-validation gives results from a large sample of logs, while the test set evaluation gives results from a smaller sample that can indicate whether the 5-fold cross-validation is biased or not.

Table 3—Accuracy metrics related to Figure 7, Figure 8, and Figure 9. For each metric, mean and sample standard deviation are calculated based on 10 training repetitions. Comparable results from a reference paper are included.

Result	Unbal. acc. (%)	Bal. acc. (%)	Precision (TP/PP) (%)	Sensitivity (TP/P) (%)	Specificity (TN/N) (%)	F1 score (%)
All annular condition codes (5-fold CV)	64.0 ± 0.30	48.0 ± 0.89	—	—	—	—
All annular condition codes (test set)	61.2 ± 0.04	47.4 ± 0.15	—	—	—	—
Isolation (5-fold CV)	83.3 ± 0.15	76.4 ± 0.23	—	—	—	—
Isolation (test set)	83.8 ± 0.03	80.1 ± 0.03	—	—	—	—
Binary (5-fold CV)	92.8 ± 0.06	91.2 ± 0.13	89.9 ± 0.16	86.9 ± 0.32	95.5 ± 0.09	88.4 ± 0.12
Binary (test set)	94.5 ± 0.01	92.9 ± 0.02	92.9 + 0.02	88.8 ± 0.04	97.0 ± 0.01	90.8 ± 0.02
Bond quality (6 classes) (Viggen et al. 2021)	57.4	50.1	—	—	—	—
Hydraulic isolation (binary) (Viggen et al. 2021)	88.9	89.5	≈ 75.5	≈ 91.0	≈ 88.0	≈ 82.5

The performance is evaluated using the metrics defined in the Quantitative Metrics section above. As the ML training process is stochastic, the results can vary between training runs. To ensure representative results, Table 3 shows the mean and sample standard deviation of each numerical metric based on 10 runs. The confusion matrices shown below originate from runs whose results were representative of this mean.

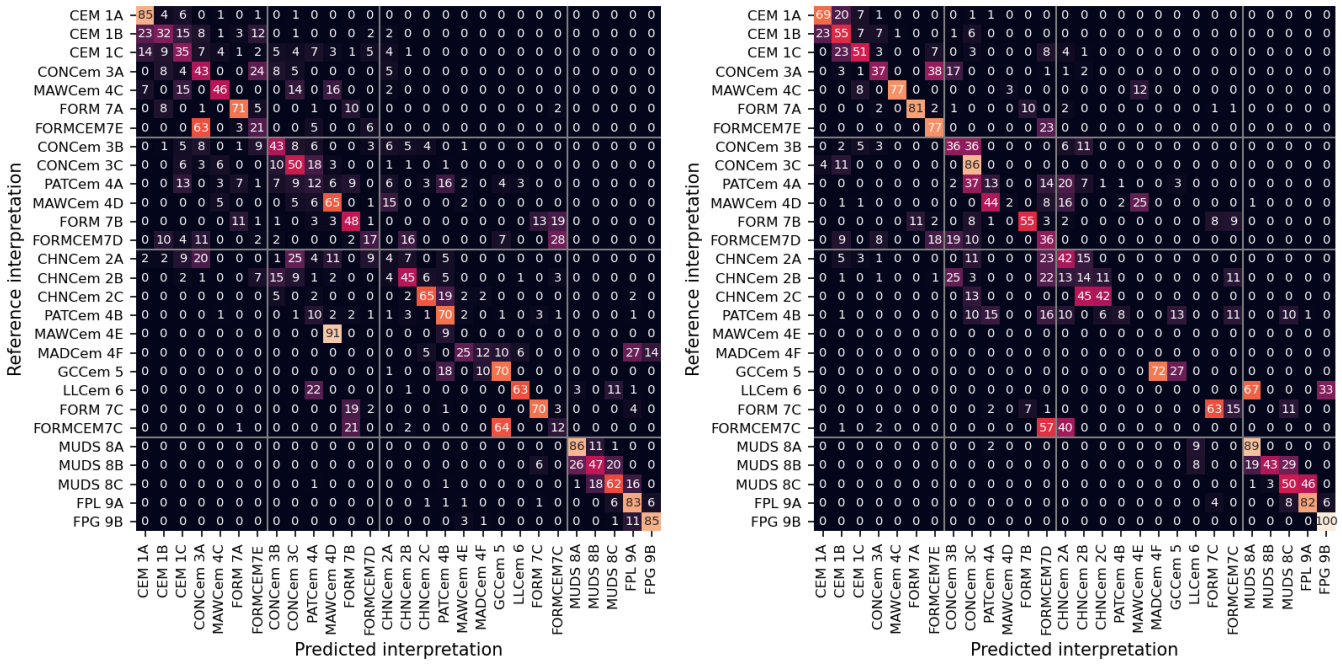

Figure 7—Balanced confusion matrices in terms of every annular condition code. The codes are sorted and gridded according to their hydraulic isolation potential. Left: 5-fold cross-validation on the training set. Right: Evaluation on the test set.

Figure 7 shows confusion matrices of the annular condition codes defined in Table 1. The confusion matrices are subdivided into grids, where annular condition codes of the same hydraulic isolation potential are grouped. Many annular condition codes are predicted well by the ML classifier, in particular those of high and no isolation potential.

However, some annular condition codes are poorly predicted. Some of these annular condition codes, such as MAWCem and FORMCEM, represent rare conditions with a low prevalence in the dataset (see Table 2). Hence, the classifier has few examples to learn from during training, as well as few cases to test its performance against. For the evaluation on the test set, the MAWCem 4E and 4F classes are missing from the confusion matrix, as these annular condition codes are not present in the test set.

Figure 7

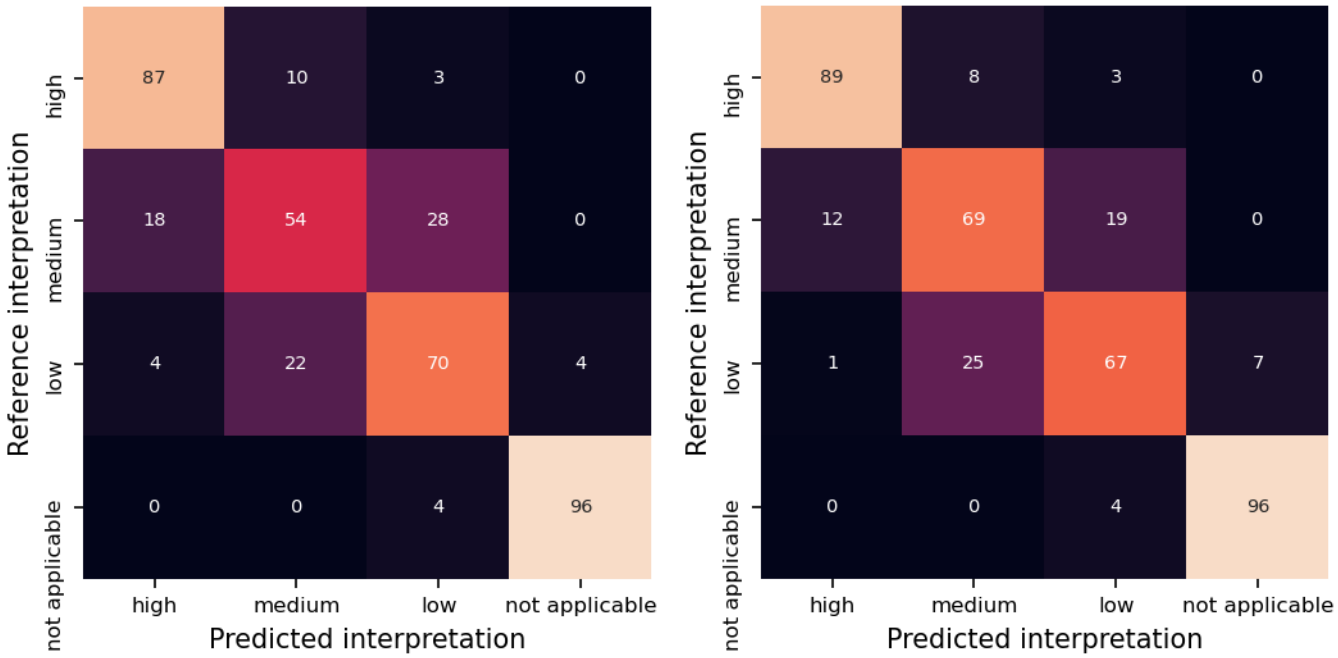

Figure 8—Balanced confusion matrices in terms of isolating potential. Left: 5-fold cross-validation on the training set. Right: Evaluation on the test set.

Figure 7 shows that when annular condition codes are confused with each other, they are largely confused with other annular condition codes of the same isolating potential. One can investigate this more directly by pooling annular condition codes according to their isolating potential. Figure 8 shows the resulting confusion matrices. The classifier is very good at identifying high isolation potential, as well as the more-or-less-free-pipe conditions identified as "not applicable" isolating potential.

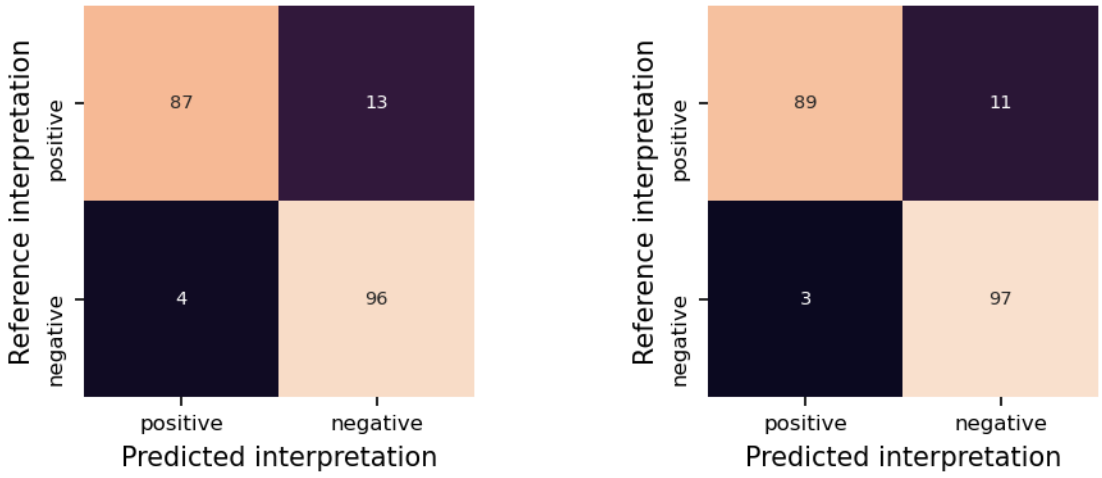

Figure 9—Balanced confusion matrices in terms of a binary classification problem on annular condition codes representing high isolating potential ("positive") versus other annular condition codes ("negative"). Left: 5-fold cross-validation on the training set. Right: Evaluation on the test set.

Going a step further, the results can be turned into a binary classification problem by pooling annular condition codes representing high isolating potential into the "positive" group, and all other annular condition codes in the "negative" group. Figure 9 shows the resulting confusion matrices. As mentioned

in the introduction, it is important to avoid false negatives and false positives, and the classifier is quite proficient at avoiding these issues. The false negative rate is only around 12%, while the false positive rate is as low as around 4%. Hence, the chance of the classifier misidentifying a non-isolating interval as isolating is low.

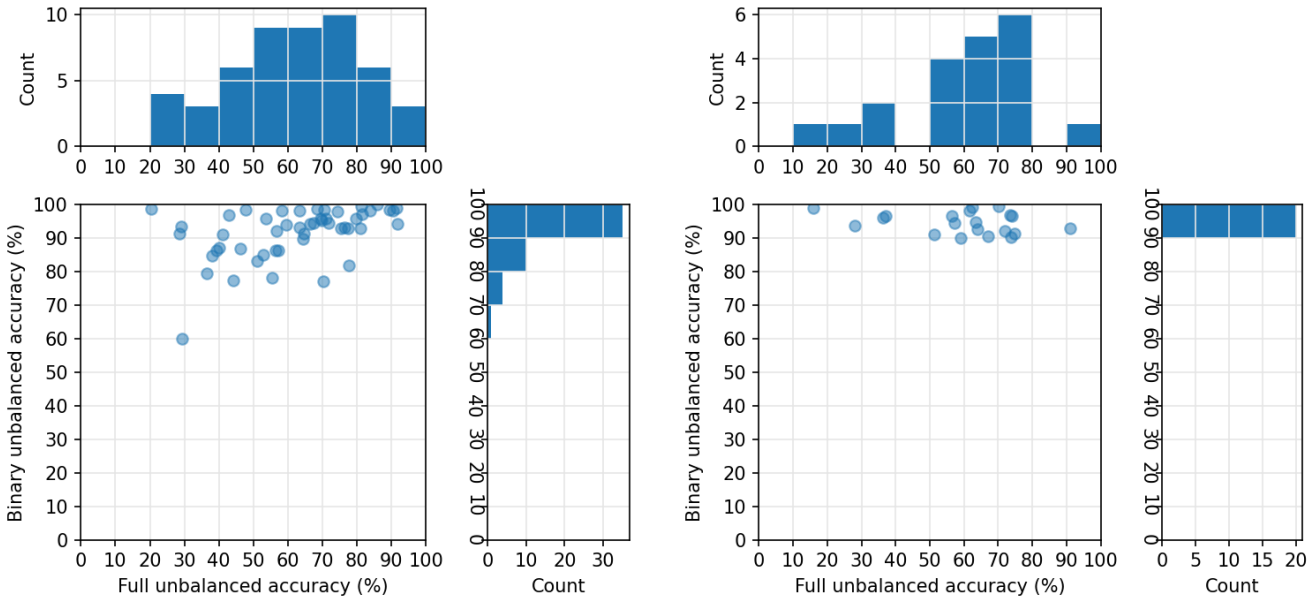

Figure 10—Joint plots of accuracy on a log-by-log basis, with cross-plots of different accuracy types and marginal histograms for each type. Left: 5-fold cross-validation on the training set. Right: Evaluation on the test set.

The results shown so far are based on either the entire training set or the entire test set. It is also relevant how much the results can vary from log to log. Figure 10 cross-plots the full accuracy (treating every code separately as in Figure 7) and the binary accuracy (as in Figure 9) for every log in the training and test sets, with marginal histograms for each type of accuracy. We see that the spread in full accuracy is much larger than the spread in binary accuracy. In the test set, every log has a binary accuracy above 90%. Furthermore, the two types of accuracy are not strongly correlated; a low full accuracy for a log does not imply a low binary accuracy or vice versa.

Discussion

Performance

The full confusion matrix in Figure 7 implies that the ML often confuses certain neighboring annular condition codes that we know are inherently difficult to be perfectly consistent with. The most straightforward examples are the MUDS 8A/8B/8C codes, which differ only in the degree of packing density of the settled mud solids. Figure 7 shows that MUDS 8B, in particular, is frequently confused with the MUDS 8A and MUDS 8C codes.

As another example, consider the classes describing well-bonded cement, CEM 1A/1B/1C. While the classifier identifies the homogeneous cement of CEM 1A well, the *heterogeneous* cement of CEM 1B is more difficult to separate from CEM 1A, both for experts and the classifier, as there is no clear-cut point of transition between homogeneous and heterogeneous cement. And the even more heterogeneous CEM 1C class, which can include small liquid pockets or channels that are unlikely to affect the hydraulic isolation, is also particularly challenging. With a more detailed zonation, the well-cemented parts of a CEM 1C zone could be identified as CEM 1A or CEM 1B zones, while the liquid pockets and channels could be identified as PATCem or CHNCem zones. It is expected that the classifier's performance suffers due to such annular condition codes, due to being trained on examples with some internal inconsistency, as well as being evaluated against similarly inconsistent data. Even so, while this inconsistency leads to a

reduction in *quantitative* performance, it might not be felt by cased hole logging interpreters as a reduction in *qualitative* performance, as they have experience dealing with such uncertainties.

Compared with a previously published work with a comparable approach whose performance was quantified on a large dataset (Viggen et al. 2021), Table 3 shows that the quantitative results in the current work are very good. This previous work interpreted logs in terms of two separate parameters: Bond quality, with 6 ordinal classes, and hydraulic isolation, with 2 classes. For bond quality, one can compare the results with those of the current full classification problem, which is much more difficult, having 28 classes instead of 6. Despite this, these current results are almost as good or better than previous bond quality results. For hydraulic isolation, one can directly compare the results with the current binary results to find that the previous results are outperformed by the current results on all but one metric.

Success Factors

Why, specifically, is the approach taken here better than that of previous work? Some success factors can be identified:

- The interpretation schema is different; while the previous work used an inherently subjective Likert scale-like interpretation schema for bond quality, the current work uses the more specific and objective interpretation schema shown in Table 1, which is likely to lead to more consistent expert interpretations
- The datasets in the current work have undergone a QC process to further improve their internal consistency
- This work has applied several other improvements suggested but not implemented by Viggen et al. (2021) to improve the feature set, for example:
    - Calculating features from both short- and longer-range context intervals
    - Including features based on metadata such as the theoretical top of cement, the time between cementing and logging, and others
- The decision to use a simpler classifier together with a more comprehensive feature engineering step instead of using a convolutional neural network approach has proven successful
- The work has been performed by a cross-functional team using an agile project methodology, and this has been very efficient and effective, despite the pandemic

Limitations

The current dataset only contains data from one service company. While various service companies can perform cement logs providing similar CBL and ultrasonic pulse-echo data, the lack of standardization in how log data should be stored and represented means that the data provided by different vendors must be handled in different ways when preparing the logs for machine learning. This handling has currently not yet been implemented for more than one service company; in the near future, the tool and the dataset will be extended to data from multiple service companies.

As mentioned in the Data Selection section, some logs are afflicted by stripes of spurious low impedance at the bottom of the casing due to mud segregation. As the assisted cement log interpretation tool takes the input log data at face value (apart from the outlier removal described in the Data Preprocessing section), it would regard these stripes as actual low-impedance channels behind the casing, which could lead to mispredictions. While one service company provides a processing algorithm that accounts for this effect of mud segregation (Thierry et al. 2016, Klieber and Lemarenko 2016), this reprocessing can only be performed on recent logs from this service company. Logs where such stripes are present and where reprocessing is not possible should not form part of the training or test sets. Furthermore, when the interpretation tool is applied to such logs, it would produce questionable predictions in the intervals where these stripes are present.

The current tool requires using the interpretation schema shown in Table 1. Consequently, the dataset must consist solely of cement logs interpreted according to this schema. While it would be relatively

straightforward to adapt the tool to use a different interpretation schema, this adaptation would also require substituting the dataset with a new dataset interpreted according to the new schema.

Furthermore, the current assisted interpretation tool can only interpret logs recorded with the same logging tools as used for the training set, or logging tools that are interchangeable with those used for the training set. As described in the Feature Extraction section, the interpretation tool is currently based on CBL curves and acoustic impedance images, which are available for all logs in the training set. Using the interpretation tool on a log with, say, an impedance image but no CBL curve would require either imputing an artificial CBL curve based on the impedance, or using an ML classifier trained without using CBL. Making use of data from other kinds of logging tools that are not present in the training dataset would require assembling a new dataset of logs produced with the desired combination of logging tools, defining features to be extracted from the new logging tools, and retraining the ML classifier on this dataset using these features.

The current interpretation tool is also limited to interpreting logs from the same kind of wells as in the current training set. For example, the logs in the training set are recorded in wells cemented with conventional cements. If the trained ML classifier were applied to a log from a well cemented with lightweight cement, whose acoustic impedance is lower than that of conventional cements, its predictions would not be reliable. Viggen et al. (2021) discuss this issue further.

Conclusions

1. This work presents a tool for assisted cement log interpretation based on supervised machine learning, where a dataset of previously interpreted acoustic cement log data is used for training the tool and evaluating its performance.
2. The implemented tool can produce an automatic interpretation of a well log to be used as a basis for an expert interpretation and is currently being beta tested by cased hole logging domain experts as an integrated part of their workflow.
3. New logs can be continuously added to the training dataset, continuously improving the performance of the tool.
4. The current work represents a significant improvement over previously published work whose performance has been quantified on a large dataset.
5. A major success factor is the usage of a more specific and objective interpretation schema that reduces the effect of subjectivity, as well as thorough quality control of the dataset.
6. The source code of this tool is intended to be shared as an open-source library. This will allow a broader usage in the industry and enable the operators and vendors to perform more consistent cement log interpretations going forward.

Acknowledgments

The authors would like to thank the project team members Ioan-Alexandru Merciu, Torgeir Haaland, and Simen Kværneng Kaspersen for their help with the design and implementation of the system. Furthermore, the authors are grateful for assistance and comments from Sally Serenyi-Smits, Selma Chikh, Kevin Constable, and the rest of the Cased-hole logging Group in Equinor. In addition, the authors appreciate the involvement and assistance of Thorvald Johannessen, Pål Viggo Hemmingsen, Therese Scheldt, Johanne Fonnes, Anne Butler Wang, Tore Tjomsland, Lilia Graciela Vazques De Holm, Lindsay Hamminga, Sarah Magdalena Birkeland, and Halvor Kjørholt. Erlend M. Viggen's contribution to the review process of the paper was supported by the Research Council of Norway under grant no. 237887.

Finally, the project would like to thank the Heidrun owners ConocoPhillips, Equinor, Petoro and Vår Energy for allowing us to use cement log examples from Heidrun as illustrations in this paper.

References

Allouche, M., Guillot, D., Hayman, A. J., Butsch, R. J., and Morris, C. W. 2006. Cement Job Evaluation. In *Well Cementing*, ed. E. B. Nelson and D. Guillot. Chap. 13, 459–501. Sugar Land, TX: Schlumberger.

API RP66 V1*, Recommended Digital Log Interchange Standard (DLIS)*, first edition. 1991. Washington, DC, USA: API.

Anderson, W. L. and Walker, T. 1961. Research Predicts Improved Cement Bond Evaluations with Acoustic Logs. *J Pet Technol* **13** (11): 1093–1097. SPE-196-PA. https://doi.org/10.2118/196-PA.

Chollet, F. & others, 2015. Keras. Available at: https://github.com/fchollet/keras

Chollet, F. 2018. *Deep Learning with Python*. Shelter Island, New York, USA: Manning Publications Co.

Daccord, G., Craster, B, Ladva, H., and Jones, T. G. J. 2006. Cement-Formation Interactions. In *Well Cementing*, ed. E. B. Nelson and D. Guillot. Chap. 13, 459–501. Sugar Land, TX: Schlumberger

Dlisio, Version 0.3.5. 2021. Oslo, Norway: Equinor. https://github.com/equinor/dlisio.

Grosmangin, M., Kokesh, F. P., and Majani, P. 1961. A Sonic Method for Analyzing the Quality of Cementation of Borehole Casings. *J Pet Technol* **13** (2): 165–171. SPE-1512-G-PA. https://doi.org/10.2118/1512-G-PA.

Hayman, A., Hutin, R., and Wright, P. 1991. High-Resolution Cementation and Corrosion Imaging by Ultrasound. Paper presented at the SPWLA 32nd Annual Logging Symposium, Midland, Texas, USA, 16–19 June. SPWLA-1991-KK.

Issabekov, Y., Kalyanaraman, R. S., Ben Haoua, T. et al. 2017. Application of New Interpretation Workflows That Improve Cement Evaluation in Presence of Microannulus. Paper presented at the SPE Annual Technical Conference and Exhibition, San Antonio, Texas, USA, 9–11 October. SPE-187461-MS. https://doi.org/10.2118/187461-MS.

Jutten, J. J. and Hayman, A. J. 1993. Microannulus Effect on Cementation Logs: Experiments and Case Histories. Paper presented at the SPE Asia Pacific Oil and Gas Conference, Singapore, 8–10 February. SPE-25377-MS. https://doi.org/10.2118/25377-MS.

Kalyanraman, R. S., van Kuijk, R., and Hori, H. 2017. Making Sense of Why Sometimes Logs Do Not See Cement in the Annulus. Paper presented at the SPE Western Regional Meeting, Bakersfield, California, USA, 23–27 April. SPE-185731-MS. https://doi.org/10.2118/185731-MS.

Kalyanraman, R. S, Chen, X, Wu, P.-Y. et al. 2021. Autonomous Interpretation Methods of Ultrasonic Data Through Machine Learning Facilitates Novel and Reliable Casing Annulus Characterization. Paper presented at the SPE/IADC International Drilling Conference and Exhibition, Online, 8–12 March. SPE/IADC-204078-MS. https://doi.org/10.2118/204078-MS

Klieber, C. and Lemarenko, M. 2016. A Calibration-Free Inversion Algorithm for Evaluating Cement Quality Behind Highly Contrasting Steel Pipe. Presented at the IEEE International Ultrasonics Symposium, Tours, France, September 18–21. https://doi.org/10.1109/ULTSYM.2016.7728565.

Kotsiantis, S.B., Zaharakis, I.D. & Pintelas, P.E. 2006. Machine learning: a review of classification and combining techniques. *Artif Intell* **26** (3): 159–190. https://doi.org/10.1007/s10462-007-9052-3

Krizhevsky, A., Sutskever, I. and Hinton, G.E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, **60** (6): 84–90. https://doi.org/10.1145/3065386

Montgomery, C. T. 2006. Implications of Cementing for Well Production and Performance. In *Well Cementing*, ed. E. B. Nelson and D. Guillot. Chap. 13, 459–501. Sugar Land, TX: Schlumberger.

NORSOK D-010*, Well integrity in drilling and well operations*, fifth edition. 2021. Oslo, Norway: Standards Norway.

Pardue, G. H., Morris, R. L., Gollwitzer, L. H. et al. 1963. Cement Bond Log: A Study of Cement and Casing Variables. *J Petrol Technol* **15** (5): 545–555. SPE-453-PA. https://doi.org/10.2118/453-PA.

Pedregosa, F., Varoquaux, G., Gramfort, A. et al. 2011. scikit-learn: Machine Learning in Python. J Mach Learn Res **12**: 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html.

Piot, B and Cuvillier, G. 2006. Primary Cementing Techniques. In *Well Cementing*, ed. E. B. Nelson and D. Guillot. Chap. 13, 459–501. Sugar Land, TX: Schlumberger.

Popović, Z. B. and Thomas, J. D. 2017. Assessing Observer Variability: A User's Guide. *Cardiovasc Diagn Ther*. **7** (3): 317–324. https://doi.org/10.21037/cdt.2017.03.12.

Press, W. H., Teukolsky, S. A., Wetterling, W. T. et al. 2007. *Numerical Recipes: The Art of Scientific Computing*. New York, NY: Cambridge University Press.

Refaeilzadeh, P., Tang, L., Liu, H. et al. 2009. Cross-Validation. In: *Encyclopedia of Database Systems*. ed. Liu, L., Özsu, M.T., 532–538 Springer: Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_565

Reolon, D., di Maggio, F., Moriggi, S. et al. 2020. Unlocking Data Analytics for the Automatic Evaluation of Cement Bond Scenarios. Paper presented at the SPWLA 61st Annual Logging Symposium, Virtual Online Webinar, 24–29 July. SPWLA-5060. https://doi.org/10.30632/SPWLA-5060.

Ting, K.M. 2017. Confusion Matrix. In: *Encyclopedia of Machine Learning and Data Mining*. ed. Sammut, C., Webb, G.I, 260-260. Springer: Boston, MA. https://doi.org/10.1007/978-1-4899-7687-1_50

Thierry, S., Klieber, C., Lemarenko, M. et al. 2016. New-Generation Ultrasonic Measurements for Quantitative Cmenet Evaluation in Heavy Muds and Thick-Wall Casings. Paper presented at the SPE Annual Technical Conference and Exhibition, Abu Dhabi, UAE, September 26–28. SPE-181450-MS. https://doi.org/10.2118/181450-MS.

Van Kuijk, R., Zeroug, S., Froelich, B. et al. 2005. A Novel Ultrasonic Cased-Hole Imager for Enhanced Cement Evaluation. Paper presented at the International Petroleum Technology Conference, Doha, Qatar, 21–23 November. IPTC-10546-MS. https://doi.org/10.2523/IPTC-10546-MS.

Viggen, E. M., Merciu, I. A., Løvstakken, L. et al. 2020. Automatic Interpretation of Cement Evaluation Logs from Cased Boreholes Using Supervised Deep Neural Networks. J Pet Sci Eng **195**: SPE-107539-PA. https://doi.org/10.1016/j.petrol.2020.107539.

Viggen, E. M., Løvstakken, L., Måsøy, S.-E. et al. 2021. Better Automatic Interpretation of Cement Evaluation Logs through Feature Engineering. SPE J **26** (5): 2894–2913. SPE-204057-PA. https://doi.org/10.2118/204057-PA.

Voleti, D. K., Reddicharla, N., Guntupalli, S. et al. 2020. Smart Way for Consistent Cement Bond Evaluation and Reducing Human Bias Using Machine Learning. Paper presented at the Abu Dhabi International Petroleum Exhibition & Conference, Abu Dhabi, UAE, 9-12 November. SPE-202742-MS. https://doi.org/10.2118/202742-MS.

Webb, G. I. 2017. Overfitting. In *Encyclopedia of Machine Learning and Data Mining*, ed. C. Sammut and G. I. Webb. Boston, MA: Springer. https://doi.org/10.1007/978-0-387-30164-8_623

Figures