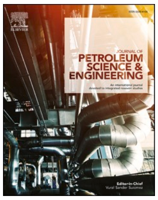
Contents lists available at ScienceDirect

Journal of Petroleum Science and Engineering

journal homepage: http://www.elsevier.com/locate/petrol



Automatic interpretation of cement evaluation logs from cased boreholes using supervised deep neural networks

Erlend Magnus Viggen[a,*], Ioan Alexandru Merciu[b], Lasse Løvstakken[a], Svein-Erik Måsøy[a]

[a] *Centre for Innovative Ultrasound Solutions, Dept. of Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, Norway*
[b] *ST DWT Well Construction, Equinor ASA, Trondheim, Norway*

ARTICLE INFO

Keywords:
Well log interpretation
Machine learning
Deep neural networks
Cement evaluation
Well integrity
Well logging

ABSTRACT

The integrity of cement in cased boreholes is typically evaluated using well logging. However, well logging results are complex and can be ambiguous, and decisions associated with significant risks may be taken based on their interpretation. Cement evaluation logs must therefore be interpreted by trained professionals. To aid these interpreters, we propose a system for automatically interpreting cement evaluation logs, which they can use as a basis for their own interpretation. This system is based on deep convolutional neural networks, which we train in a supervised manner using a dataset of around 60 km of interpreted well log data. Thus, the networks learn the connections between data and interpretations during training. More specifically, the task of the networks is to classify the bond quality (among 6 ordinal classes) and the hydraulic isolation (2 classes) in each 1m depth segment of each well based on the surrounding 13 m of well log data. We quantify the networks' performance by comparing over all segments how well the networks' interpretations of unseen data match the reference interpretations. For bond quality, the networks' interpretation exactly matches the reference 51.6% of the time and is off by no more than one class 88.5% of the time. For hydraulic isolation, the interpretations match the reference 86.7% of the time. For comparison, a random-guess baseline gives matches of 16.7%, 44.4%, and 50%, respectively. We also compare with how well human reinterpretations of the log data match the reference interpretations, finding that the networks match the reference somewhat better. This may be linked to the networks learning and sharing the biases of the team behind the reference interpretations. An analysis of the results indicates that the subjectivity inherent in the interpretation process (and thereby in the reference interpretations we used for training and testing) is the main reason why we were not able to achieve an even better match between the networks and the reference.

1. Introduction

Cementing is a very common operation carried out during the construction phase of the majority of oil wells. The idea of cementing operations can be traced back to 1859 and 1871, with the first cement operation executed in 1883 by Hardison & Stewart Oil Company (Mau and Edmundson, 2015; Hill, 1871). Cementing operations have two main objectives. The first objective is to provide well integrity by controlling flow in the well through hydraulic isolation between different zones in the wellbore. Thus, successful cementing prevents fluids from geological formations flowing into other geological zones or to the surface. The second objective is to provide support for the casing.

To ensure that a cementing job was successful, we must test the cement. Older cementing jobs may also be tested again to ensure that they still hold, e.g. as a step in cost-effective plug & abandonment operations (Vrålstad et al., 2019). To date, the only method that can confirm zonal isolation with certainty is a pressure test. However, pressure tests may be economically unfeasible, and field experience shows that they in some cases risk causing damage to the cement. Therefore, companies typically evaluate cement through well logging, where measurement tools are lowered into the casing string to check the presence and quality of the cement on its outside.

Since Grosmangin et al. (1961) and Anderson and Walker (1961) published the first effective cement evaluation method, which was based on sound waves, further tool development has been somewhat slow. Even today's techniques cannot unambiguously verify the presence of

[*] Corresponding author.
E-mail addresses: erlend.viggen@ntnu.no (E.M. Viggen), iom@equinor.com (I.A. Merciu), lasse.lovstakken@ntnu.no (L. Løvstakken), svein-erik.masoy@ntnu.no (S.-E. Måsøy).

https://doi.org/10.1016/j.petrol.2020.107539
Received 2 March 2020; Accepted 12 June 2020
Available online 2 July 2020

isolating cement. Among the existing methods, however, acoustic logging techniques are the most common and efficient according to Allouche et al. (2006). These comprise sonic and ultrasonic techniques (Grosmangin et al., 1961; Hayman et al., 1991; van Kuijk et al., 2005). The data recorded by acoustic tools may be processed to get estimates of various parameters that describe the state of well components such as the casing and the cement. The well's potential for hydraulic isolation may then be interpreted from these results.

However, this interpretation is a complex task with associated risks (Benge, 2014; Kyi and Wang, 2015), and it must therefore be carried out by trained professionals. They use their understanding to integrate the various log results and their knowledge about the well to produce an evaluation of the cement status. We describe this process further in Sec. 2.1.

This task is performed under time pressure, as further well development may hinge on the evaluation results. As Belozerov et al. (2018) also point out, the process of well interpretation is complex, time consuming, and quite subjective as well: Although no studies on this have yet been published, oil companies are well aware that different interpreters can reach different conclusions from the same data. We consider this subjectivity further in Sec. 2.2.

In the work presented here, we created and tested an automatic system to generate well log interpretations from log data. Our goal is to help interpreters offset the aforementioned problems by using the system's automatic interpretations as a baseline for manual interpretation. As the system is trained on previous manual interpretations, it may thus help keep future interpretations more consistent. The system would also speed up the interpretation process if human interpreters only need to correct its interpretations where it fails. While humans may always need to be part of the loop when important decisions are made based on well logs, such a system could also provide quick-look interpretations for less important wells. We give an overview of our approach to creating such an automatic system in Sec. 2.3.

2. Background

2.1. Manual well interpretation

When interpreting a cement evaluation log, the well integrity interpreter looks at a plot of various log results against depth. (Fig. 2 shows three examples of such plots.). The task is to explain the results by partitioning the well into zones, or intervals, answering two main questions for each of them: ‘What is the bonding between annular solids and the casing in this zone?” and “What is the zone's potential for hydraulic isolation?’ The integrity interpreter must have access to the log results, either physically on paper or digitally through specialised software that can read well, plot, and process log data. The interpreter must also have access to the well history, which provides context to the well log (Benge, 2014). Table 1 shows part of an interpretation resulting from such a task.

Table 1
Extract of the official interpretation of the well Volve 15/9-F-9 after a log operation in June 2009 (Equinor, 2018).

Interval top [mMD]	Interval bottom [mMD]	Cement or formation bond quality	Potential for hydraulic isolation
150	178	Moderate to poor	Low
178	188	Good to moderate	Medium
188	195	Moderate	Medium
195	201	Good to moderate	Medium
201	221	Poor	Low
221	332	Poor	Low
332	341	Moderate to poor	Low
341	440	Moderate to poor	Low
440	451	Moderate	Medium

Fundamentally, interpreting cement logs is similar to interpreting open-hole logs. In both cases, the interpreter must understand the physics and processing underlying different measurements, and the limits thereof, so that they can weigh different measurements against each other (Kyi and Wang, 2015). The interpreter must therefore be trained in the art of borehole logging and imaging in order to reduce interpretation error. Furthermore, the interpreter must be familiar with the decisions that hinge on the interpretation, and the potential risks therein.

Unlike open-hole log interpretation, where nomenclature, rules, and documentation is generally agreed upon, integrity interpreters lack a general recipe for interpretation. Another challenge of cased-hole interpretation relates to the difficulty of correctly interpreting vertical features. While interpreting azimuthal heterogeneities in cased holes tends to be easier, vertical heterogeneities such as fluid channels raise questions such as whether the channels are connected, how long the channels are, and whether the integrity of that zone is compromised. The interpreter must also consider possible reasons why the vertical features exist in the first place. A problem of comparable complexity from open-hole logging is the interpretation of drilling-induced fractures in images from spiral holes.

2.1.1. Interpretation of acoustic logs

Over the past decades, well integrity interpreters have mainly been exposed to and familiarised with results from acoustic logging methods. These results are extracted from raw acoustic waveforms through time or frequency domain-based processing (Pardue et al., 1963; Hayman et al., 1991; Allouche et al., 2006). While laboratory tests can evaluate the integrity of cement sheaths in detail (Albawi et al., 2014), the limitations of current acoustic technologies mean that a much smaller amount of information is extracted during logging. This information can be ambiguous, and interpreters must therefore often make strong assumptions, for example that intervals with solids bonded to the outside of the casing can be interpreted as isolated.

Acoustic logging is based on acoustic waves propagating in the well. There are two subtypes: Sonic logging is lower-frequency (10–80 kHz) and nondirective, while ultrasonic logging is higher-frequency (0.1–2 MHz) and directive. Both require tool immersion in a liquid-filled environment inside the casing that is free of debris or thin oil films. They also require centralisation inside the casing. If any of these requirements is not completely fulfilled, the interpreter must take the resulting uncertainties into account while interpreting.

Sonic tools' monopole transmitters impinge omnidirectional pressure pulses onto the casing. There, they excite extensional waves (specifically, in the low-dispersive S_0 Lamb mode) travelling up and down the cases, as well as wavefields in the annulus and formation (Pardue et al., 1963; Tubman et al., 1984, 1986; Sinha and Zeroug, 1999; Wang et al., 2016). These wavefields in turn feed back into the wavefield in the fluid inside the casing. This is recorded by two hydrophones at distances of 3 ft (0.914 m) and 5 ft (1.524 m) from the transmitter. The main feature extracted from this wavefield is the cement bond log (CBL) data channel. CBL is the signal amplitude in mV of the component of the wavefield that arrives first at the closest hydrophone. This component has propagated along the casing, as this is the fastest path from transmitter to receiver. As the casing wave continuously loses energy into the materials on both sides of the casing, and the loss into bonded solids is strong, lower CBL values signify solids bonded to the outside of the casing. The CBL value can roughly be interpreted by thresholding. If the CBL is below around 10 mV, this indicates full azimuthal coverage of solids. If the CBL is around a certain high value (depending on casing size and normalisation, though typically 50–60 mV), this indicates free pipe, i.e. no solids bonded to the casing. As CBL varies with depth, it is typically plotted as a curve. The full recorded waveforms for every depth can also be plotted as an image, with depth along one dimension and waveform samples along the other. This is called a variable density log (VDL), and this data channel can be used for further qualitative interpretation, as we will see in Sec. 2.1.2. Fig. 2 shows examples of CBL curves and VDL images for three sections in a well.

Unlike sonic tools, which can only sample the well in depth, ultrasonic tools can also sample the well azimuthally. Pulse-echo tools impinge ultrasonic pulses onto the casing at normal incidence and record the echo. Pitch-catch tools use a transmitter to impinge pulses on the casing at oblique incidence, creating a leaky flexural Lamb wave (specifically, in the A0 mode) whose wavefield is recorded by one or more receivers. Depending on the type of tool, these recordings can be processed into data channels containing estimates of various physical quantities as a function of depth and azimuth. For example, pulse-echo tools can estimate the inner radius and thickness of the casing and the impedance of the material behind it (Zemanek et al., 1969; Froelich et al., 1982; Hayman et al., 1991; Graham et al., 1997). Pitch-catch tools can provide additional information on the material behind the casing through the attenuation of casing Lamb modes, and they may also provide information on the third interface beyond the annulus, which may be formation rock or an outer casing (van Kuijk et al., 2005; Herold et al., 2006; Morris et al., 2007). These depth-by-azimuth maps of physical quantities are typically plotted as images, as can be seen in Fig. 2.

There is some redundancy between the sonic and ultrasonic measurements. CBL indicates the degree of bonded solids outside the casing, while ultrasonic pulse-echo tools can estimate the acoustic impedance of the outside material, where high impedances indicate bonded solids. The higher the overall impedance surrounding the casing, the lower the CBL, and vice versa. This correspondence can be seen clearly in Fig. 2, where the impedance shown in the AIBK image channel is clearly correlated with the CBL values shown in the CBLF curve channel.

In general, interpreting well integrity based on acoustic logs entails evaluating the quality of the bonding of outside solids to the casing. However, the acoustic logs sometimes tell an ambiguous story, from which different interpretations can be made. Transforming this story into a correct understanding of the well status is a difficult task that may require using information drawn from other sources, such as the well development, cementing, and logging histories.

2.1.2. *Example of manual interpretation*

As an example of manual interpretation to provide further background for the rest of the paper, we show and interpret logs in three relatively straightforward well sections. These logs were recorded in a casing with an outer diameter of $9\frac{5}{8}$in (24.45 cm) and thickness of 0.539 in (1.37 cm) in the well Volve 15/9-F-11 B, and are freely available from Equinor (2018). Fig. 1 shows the wellsketch, and Fig. 2 shows the well logs for the three sections. The objective of this log operation was to evaluate cement quality in relation to requirements for production packer placement, and to identify the top of cement (ToC).

The log data was recorded using a sonic tool and pulse-echo ultrasonic tool as described in Sec. 2.1.1, both from the same service company. The data quality is good, apart from some eccentralisation of the tools. The theoretical ToC was estimated at 2670 m, adjusted to 2980 m to account for losses during cementing. This tells the interpreter that there is little chance that solids above 2670 m correspond to cement, and that section 2670–2980 m should be interpreted pragmatically.

Section 2625–2645 m: From the theoretical ToC, there is little to no chance of cement in this interval. The tool eccentering in this section is a little higher than the recommended maximum of 2% of casing outer diameter (Hayman et al., 1991), which for this casing is 0.1925 in (4.890 mm). This eccentering can also be seen from the amplitude of the reflected pulse (AWBK), which is much higher on the bottom side of the casing (image centre). There are few quality control (QC) flags in the UFLG channel, which supports that the data quality is sufficient for interpretation. The casing inner radius (IRBK) and thickness (THBK) channels show only slight ovalisation of the casing and no other detectable casing defects. As always, casing collars such as the one at 2635 m cause localised disturbances in various channels that must be disregarded, such as spuriously high acoustic impedances behind the casing (AIBK), spikes in CBL, and chevron patterns in VDL.

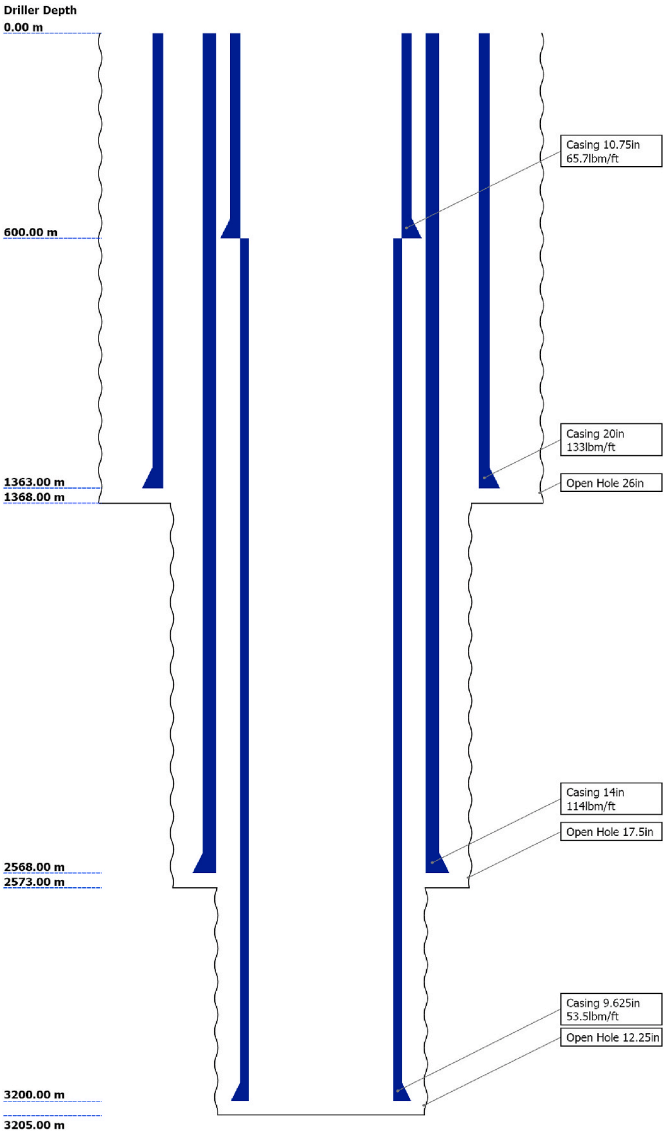


Fig. 1. Wellsketch of Volve 15/9-F-11 B, from Equinor (2018).

The impedance values are quite heterogeneous, varying from 2 to 6 MRayl. There is a large galaxy pattern at 2626–2631m, which indicates that the casing is close to the borehole wall on the bottom side (Hayman et al., 1991; Miller and Stanke, 1999). With this information in mind, the higher-impedance patterns further down are possible formation footprints, either due to casing eccentering or incipient collapse at the time of logging. The CBL generally shows high values. The VDL dimming and showing formation arrivals below 2637 m reinforce the view that formation may be touching the casing here.

Even taking eccentering-related uncertainties into account, the ultrasonic and sonic measurements in this section lead to an interpretation of free pipe in the upper half, poor bond quality in the lower half, and no potential for hydraulic isolation.

Section 2800–2820 m: From the theoretical ToC, there is a chance of cement in this section. The QC situation is similar to the previous section, except for a stronger eccentering and casing centralisers being visible on the impedance log at 2805, 2811, and 2817 m.

In this section, the impedance is even more strongly heterogeneous, with values varying from 2 to 9 Mrayl. (While the figure plots impedance

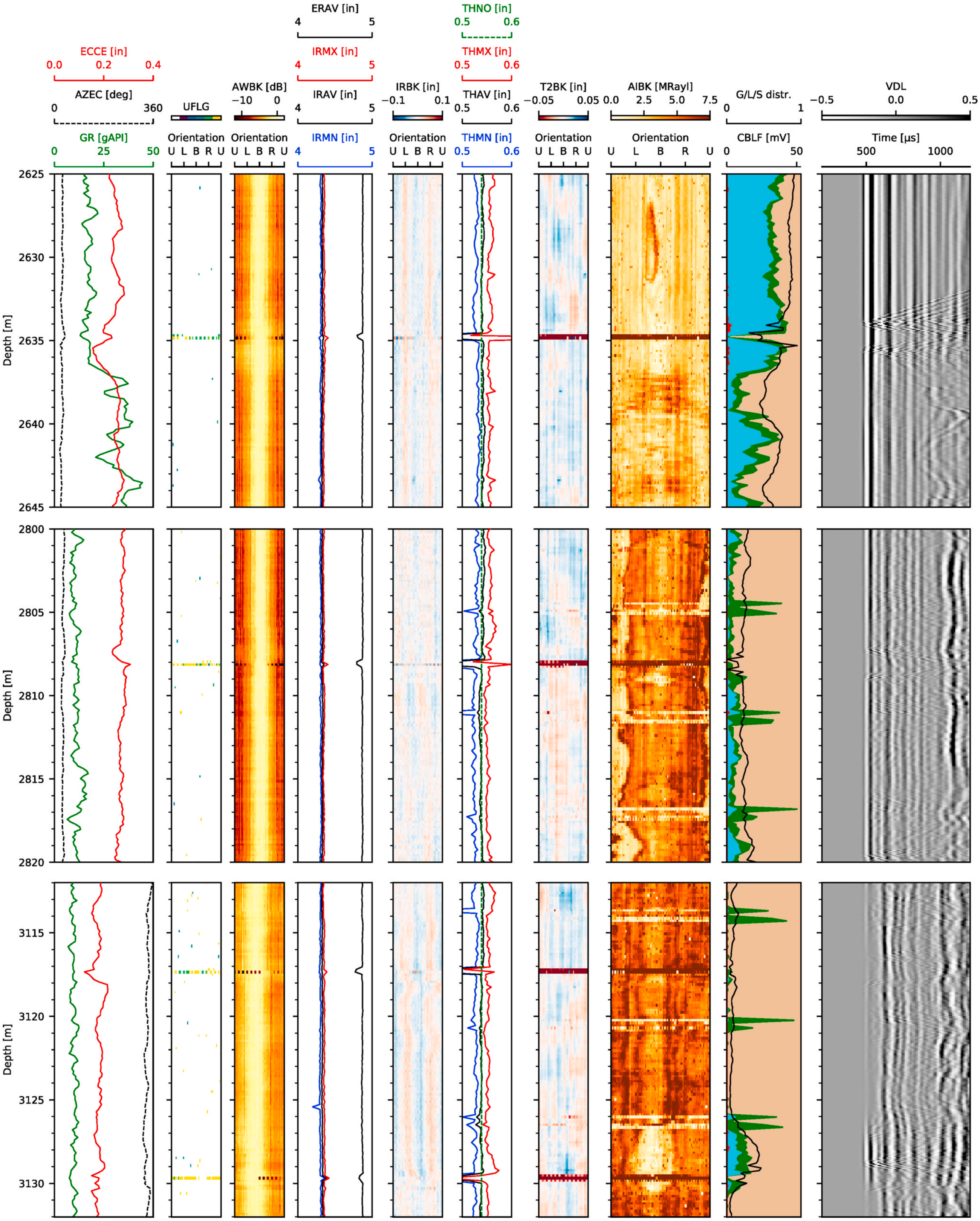


Fig. 2. Log plot of data channels from three sections from Volve 15/9-F-11 B. The plotted quantities are explained in Sec. 3.2, with the exception of IRMN and IRMX (minimum and maximum casing inner radius), ERAV (average casing outer radius), and THMN, THAV, THMX, and THNO (minimum, average, maximum, and nominal casing thickness). The G/L/S distr. column estimates the share of gas (red), liquid (blue), microdebonded solids (green) and solids (brown), mainly by thresholding impedance as described by Allouche et al. (2006). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

with a standard upper limit of 7.5 MRayl, we found this range directly from the data.) We observe a large vertical channel of low impedance at the top of the casing (image sides), which indicates fluid pockets interconnected for the length of the section. The impedance image also shows traces of possible formation footprint as higher-impedance horizontal features at 2801 m and 2815 m, though it cannot be concluded that the formation provides bonding support. The CBL value is moderate throughout the section, while the VDL shows formation arrivals throughout the section.

Even taking eccentering-related uncertainties into account, the sonic and ultrasonic measurements lead to an interpretation of moderate bond quality and low potential for hydraulic isolation due to the presence of the top channel.

Section 3112–3132 m: From the theoretical ToC, we can expect cement to exist in this section. The QC situation is similar to the previous sections, except that the eccentering is now largely within the recommended maximum.

Again, the impedance is heterogeneous, varying from 2 to 9 MRayl. The impedance is generally high and CBL generally low, which indicates good solid coverage, with the exception of a liquid pocket at 3126–3131 m. This pocket is visible on the ultrasonic, CBL, and VDL logs. The impedance also contains a heterogeneous vertical feature which may indicate two phases of solid deposition at different times. However, there is little to no chance of a liquid channel along the ultrasonic image outside the liquid patch.

The ultrasonic and sonic measurements lead to an interpretation of high bond quality and high hydraulic isolation in this section, except for the localised liquid pocket at 3126–3131 m, which should be interpreted further in the context of the entire log.

2.2. Subjectivity in interpretation

In many fields, different experts can look at the same data and come to different conclusions. For example, two medical doctors may come up with two different medical diagnoses based on the same signs and symptoms, or two psychiatrists may make different diagnoses after having listened to the same clinical interview. Two components of this variability are often considered (Popović and Thomas, 2017):

Interobserver variability, the tendency of different observers to make different judgements when interpreting the same data.

Intraobserver variability, the tendency of a single observer to make different judgements when interpreting the same data multiple times.

In well log interpretation, this variability can manifest itself in several ways. Interpreters may disagree on which label to apply to a well interval, for example if one interpreter sees inhomogeneous solids around the entire cross-section, whereas another sees a fluid channel through cement. They may also disagree on where to place the boundaries between interpreted intervals, and how fine to make their interpretations. For example, when interpreting a long stretch of patchy cement, one interpreter may use a single long interval, whereas another may use many intervals, one for each patch and one for each stretch between patches.

This variability can be partly offset by having a team of interpreters collaborating on interpretation tasks, although this further increases the time an interpretation requires. For the dataset used in this article, the collaboration approach was to have a first interpreter perform a complete interpretation of each log, and then run each interpretation through a quality control process where one or more highly experienced individuals in a team of interpreters examine and correct it. (However, while the members of a team may develop a common understanding of how to interpret integrity, this understanding may not be universal, and the team's approach may differ from that of another team.)

2.3. Automatic interpretation through supervised learning

From the preceding sections, we see that making an accurate well integrity interpretation is a very difficult problem. Another difficult problem would be to write a computer program that duplicates the decision-making process of a human interpreter. Firstly, it would require a complete understanding of that process for every case that that may come up when interpreting an integrity log. Secondly, it would require translating this process into code.

Another path towards an automatic interpretation system is machine learning, in particular supervised learning, where an algorithm is trained on already-interpreted data. A major advantage of supervised learning is that we do not need to implement the decision-making process ourselves. Instead, we show interpreted data to the supervised learning algorithm, and it learns by itself the connections between certain types of log data and their interpretations. If done correctly, the algorithm is then able to make reasonable interpretations of data that it did not see during training.

Research on machine learning on well log data is currently taking off, with many papers on the topic published at the SPWLA 60th Annual Logging Symposium in 2019 (Bennis and Torres-Verdín, 2019; Bigoni et al., 2019; Dai et al., 2019; Gupta et al., 2019; Jain et al., 2019; Li et al., 2019; Liang et al., 2019; Oruganti et al., 2019; Peyret et al., 2019; Shao et al., 2019; Wu et al., 2019). Only one paper on the topic had previously been published at the symposium (Akkurt et al., 2018).

Interesting papers on the topic have also appeared elsewhere. For example, Onalo et al. (2018) used neural networks to recreate open-hole sonic logs from other open-hole log data for cases where reliable sonic logs were not available, Belozerov et al. (2018) used neural networks to identify oil reservoirs from well log data, and Gkortsas et al. (2019) used support vector machines and neural networks to automatically identify an ultrasonic waveform feature that can give additional information on the P-wave speed of the annular material in cased boreholes.

However, using machine learning to replicate manual interpretation of cement quality is a quite difficult problem. In particular, it is difficult because machine learning typically requires a lot of data, and sufficient amounts of interpreted log data is hard to come by outside of oil companies. Additionally, the log data can be quite heterogeneous: Different log runs can use different tools whose measurements cannot be compared directly, and even the same tool can have different values of settings such as resolution across different runs.

The task of interpreting the well status based on well log channels is very similar to the general task of image classification. In both tasks, periodically sampled data is analysed in order to classify it according to what it contains. Classification of photos has been very extensively studied over the last decade, with current best approaches based on convolutional neural networks (CNNs) (Russakovsky et al., 2015; Chollet, 2018). For this reason, the work presented here is also based on CNNs. CNNs are also widely used for image classification tasks in other fields, such as medicine (Anthimopoulos et al., 2016; Cheng and Malhi, 2017; Østvik et al., 2019). In our case, however, the task is somewhat more difficult, as we do not classify the well status from single images. Rather, we must use a heterogeneous collection of image and curve data channels for our classification.

3. Data and methods

3.1. Datasets

The dataset underlying this work is the combination of two individual datasets. The first is the public Volve Data Village dataset from Equinor (2018). It contains, among a great deal of other data, interpreted cement evaluation log data from three wells in the Volve field. These were recorded from 2009 to 2016. The second dataset contains interpreted cement evaluation log data from 29 wells in another field. The data was recorded from 2009 to 2012.

In total, we have official interpretations from 54 logging operations, all from the same team of interpreters. The interpretation process is described in Sec. 2.1, and an example extract of a manual interpretation is shown in Table 1. The shortest and longest interpreted intervals in our dataset are 1m and 1783 m long respectively, and the median interval is 33.5 m long. Where log data was available, we associated each interpreted interval with one or more log data files containing the data that was interpreted. Repeat passes were included wherever visual comparison of the data showed that their calibration and alignment matched the main passes. Thus, our 54 interpretations were associated with 99 data files that together contain 62594 m of interpreted log data.

3.2. Logging tools and data channels

In our dataset, all log data files contain data from a sonic and/or ultrasonic tool, as described in Sec. 2.1.1. The sonic tool used is mainly Schlumberger's Digitizing Sonic Log Tool (DSLT), and the ultrasonic tool is mainly Schlumberger's Ultrasonic Imager Tool (USIT), which uses the pulse-echo technique described in Sec. 2.1.1.

Each tool provides many data channels containing measured or processed data. In a log file, each data channel is sampled at a constant depth resolution. Curve channels hold only a single value per depth, while image channels hold multiple values (representing data over, e.g., a set of azimuthal angles or the samples of time signals) per depth. Example data for the most important channels are shown in Fig. 2, and the channels used as input to our neural networks are:

CBLF: Sonic curve channel; corrected cement bond log.

VDL: Sonic image channel; received sonic waveforms.

ECCE/AZEC: USIT curve channels; eccentering magnitude and direction.

AWBK: USIT image channel; amplitude of returned pulse.

IRAV: USIT curve channel; average casing inner radius.

IRBK: USIT image channel; deviation of casing inner radius from IRAV.

T2BK: USIT image channel; deviation of casing thickness from its nominal value.

AIBK: USIT image channel; acoustic impedance of material behind the casing.

UFLG: USIT image channel; QC flags indicating where USIT processing fails.

UCAZ/RB: USIT curve channels; rotation of USIT images.

GR: Gamma ray curve channel; local radioactivity from formation.

For logs measured using Schlumberger's Sonic Scanner instead of the DSLT, we used its discriminated synthetic CBL (DCBL) curve channel as a drop-in replacement for CBLF.

A problem with a dataset such as ours is that it is very heterogeneous. The depth resolution of each channel can vary from log run to log run. Some channels are also missing in some log files, typically (but not always) because the tool producing it was not present on the tool string. One logging operation also had to be excluded as it only used a slim sonic tool whose CBLF values were not properly normalised.

3.3. Data extraction

3.3.1. Well log interpretations

The interpretations in the dataset originally came in the form of tables in log report documents. As these report tables did not all use the same set of interpretation parameters, we defined a common table format that we could manually copy the report tables into. For example, data from report table columns titled 'Hydraulic isolation' (with yes/no labels), 'Probability of hydraulic isolation' (with low/medium/high and intermediate labels), and 'Isolating potential based on cement or formation quality' (with low/medium/high labels) were all gathered under a column title 'Hydraulic isolation'.

Furthermore, while the report tables' labelling was somewhat freeform, we could translate the labels into six bond quality (BQ) classes $\mathrm{C}^{\mathrm{BQ}} = \{C_0^{\mathrm{BQ}}, \dots, C_5^{\mathrm{BQ}}\}$ and two hydraulic isolation (HI) classes $\mathrm{C}^{\mathrm{HI}} = \{C_0^{\mathrm{HI}}, C_1^{\mathrm{HI}}\}$, shown by name in the header of Table 2. The BQ classes are ordered, i.e., they form an ordinal scale. For HI, original 'High' and 'Yes' labels were translated to 'Yes', and others were translated to 'No or uncertain'. (We did not distribute HI more finely as the resulting classes would have become very unbalanced due to the scarcity of intermediate labels.) We excluded log reports that did not provide interpretations of both BQ and HI.

3.3.2. Well log data

The well log data files were provided in the archaic yet widely-used DLIS file format (API, 1991). As DLIS files would be too slow to continuously read from during training, we mirrored the files' contents to an HDF5-based format using the Java library Log I/O, developed by Petroware. During training, we can read directly and efficiently from these HDF5 files.

A data channel may be provided at different depth resolutions in different log data files. However, our neural networks needs to have its input data at a consistent resolution. For that reason, we defined a target resolution for each data channel. The data of each channel then has to be transformed from its original resolution to the target resolution in as fast a manner as possible, so that this transformation will not slow down the neural network training. Where the original resolution is higher and an integer multiple of the target resolution, we stride through the data channels. Elsewhere, we use nearest-neighbour interpolation in depth. Unlike other basic interpolation methods, nearest-neighbour interpolation naturally handles NaN values, which data channels frequently use to represent missing data.

3.4. Data organisation

We split all of the interpreted intervals into depth segments of 1m length. We then tied each segment to the 13m interval of log data surrounding it, as shown in Fig. 3. This ensures that the automatic interpretation of each segment is supported by log data also outside the segment. This choice for example makes it possible to differentiate between fluid patches and fluid channels when interpreting HI. We chose the interval length as 13 m to ensure that each log data interval contains at least one casing collar, with casing joints being around 12 m long.

In total, we have 58781 samples in the dataset, where each sample i represents the interpretation labels $y_i^{\mathrm{BQ}}, y_i^{\mathrm{HI}}$ of a 1 m segment and data X_i from the corresponding 13 m log data interval. We partitioned these samples into six folds. To ensure that no two folds contain information from the same well section, we ensured that all logs with the same well and casing size are placed in the same fold. We optimised this partitioning using simulated annealing to ensure that the folds were of approximately equal size, with approximately the same distribution of classes for both BQ and HI. Table 2 shows the distribution of each fold.

3.5. Accuracy and metrics

In well log interpretation, there is no ground truth available. The well is far underground and inaccessible, and its true status can only be determined through cement core retrieval (Crow et al., 2009), an expensive and destructive and therefore very rare operation. Therefore, the task of our supervised learning system cannot be to come up with a 'true' interpretation, but rather to reproduce expert interpretations as well as possible. The correspondence between two such interpretations can be quantified using various accuracy metrics by treating one as a reference and the other as a prediction. The resulting accuracies then quantify how well the predicted interpretation matches the reference interpretation.

The most basic accuracy metric we use is the *precise accuracy*, which is the proportion of N samples where the predicted BQ or HI label $\hat{y}_i$ equals the reference label y_i:

Table 2
Distribution of the samples in each of the six folds among classes of bond quality and hydraulic isolation.

		Bond quality (BQ)						Hydraulic isolation (HI)	
Fold	**No. of samples**	Good	Moderate to good	Moderate	Poor to moderate	Poor	Free pipe	Yes	No or uncertain
0	9309	2270	423	957	983	1877	2799	2830	6479
1	9922	2428	414	930	1010	1968	3172	2881	7041
2	9652	2285	436	884	1020	2127	2900	2520	7132
3	10 104	2226	422	841	807	2572	3236	2787	7317
4	9954	1922	743	964	872	2462	2991	3124	6830
5	9676	2085	398	873	954	2110	3256	2789	6887
Total	58 617	13 216	2836	5449	5646	13 116	18 354	16 931	41 686

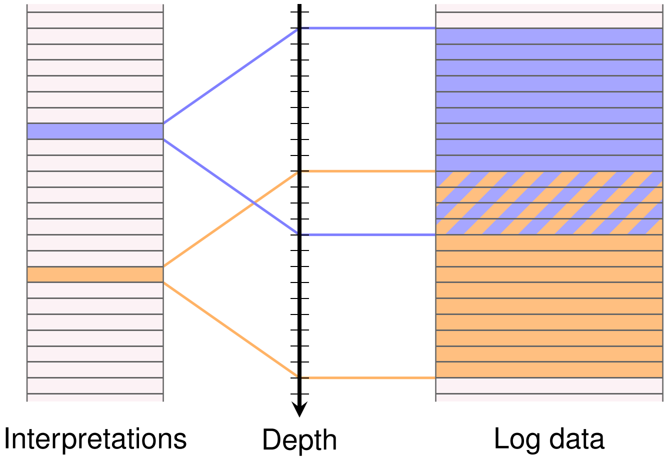


Fig. 3. Each 1 m interpreted well segment is tied to a 13 m log data interval surrounding it. Segments that are close together, such as the two shown here, will use partly overlapping intervals of log data.

$$\mathrm{UPA}(y,\hat{y}) = \frac{1}{N}\sum_{i=0}^{N-1} 1(\hat{y}_i = y_i)\ . \tag{1a}$$

Here, $1(x)$ is an indicator function which equals 1 if its argument x is true and 0 otherwise. For BQ classification, we also use the *adjacent accuracy*

$$\mathrm{UAA}(y,\hat{y}) = \frac{1}{N}\sum_{i=0}^{N-1} 1\left(\hat{y}_i^{\mathrm{BQ}} \approx y_i^{\mathrm{BQ}}\right)\ , \tag{1b}$$

where $\hat{y}_i^{\mathrm{BQ}} \approx y_i^{\mathrm{BQ}}$ holds true if the predicted label $\hat{y}_i^{\mathrm{BQ}}$ is off by no more than one class from the reference label y_i^{BQ}. The adjacent accuracy is therefore the proportion of samples for which this holds. The idea of adjacent accuracy is that, for example, predictions of 'Moderate to good' or 'Poor to moderate' may still be close enough to a reference label of 'Moderate' to be useful. (For HI, using adjacent accuracy does not make sense as there are only two classes; the adjacent accuracy would have been 100% in every case.)

We also report these accuracies in *balanced* form, using the same definitions as the scikit-learn library by Pedregosa et al. (2011). Balanced accuracy compensates for the imbalance of classes in the dataset, seen in Table 2, by weighting each sample as $w_i = 1/\left[\sum_j 1(y_j = y_i)\right]$, i.e. inversely proportional to its class' prevalence in the dataset:

$$\mathrm{BPA}(y,\hat{y}) = \frac{1}{\sum_{i=0}^{N-1} w_i}\sum_{i=0}^{N-1} 1(\hat{y}_i = y_i)w_i \tag{1c}$$

$$\mathrm{BAA}(y,\hat{y}) = \frac{1}{\sum_{i=0}^{N-1} w_i}\sum_{i=0}^{N-1} 1\left(\hat{y}_i^{\mathrm{BQ}} \approx y_i^{\mathrm{BQ}}\right)w_i\ . \tag{1d}$$

These metrics ensure that the results of every class has the same weight. The *unbalanced* accuracy metrics in (1a) and (1b) do not, and therefore emphasise the more common classes, namely 'Good', 'Poor', and 'Free pipe' for BQ and 'No or uncertain' for HI.

In summary, we use four accuracy metrics: Unbalanced precise accuracy (UPA), balanced precise accuracy (BPA), unbalanced adjacent accuracy (UAA), and balanced adjacent accuracy (BAA).

To show the distribution of labels in more detail, we also use balanced confusion matrices, which show the joint distribution of the reference labels y_i and the predicted labels $\hat{y}_i$:

$$\mathrm{CM}_{k,l}(y,\hat{y}) = \frac{1}{\sum_{i=0}^{N-1} 1(y_i = C_k)}\sum_{i=0}^{N-1} 1(y_i = C_k, \hat{y}_i = C_l). \tag{1e}$$

The vertical axis indexes the classes C_k for the reference, while the horizontal axis indexes the classes C_l for the prediction. Balancing in this way, all matrix rows sum up to 100%.

Now, what kind of accuracy can we expect? As a lowest baseline for our classification problems, we have a classifier that simply guesses randomly. For BQ, this would give a precise accuracy of 16.7% and an adjacent accuracy of 44.4%. For HI, the corresponding precise accuracy is 50%. On the high side, we cannot expect accuracies close to 100%; due to the subjectivity discussed in Sec. 2.2, this would be unattainable even by human interpreters. Instead, there must be an upper accuracy threshold, related to the subjectivity inherent in the manual labels on which we train and test.

3.6. Manual reinterpretation

To shed some light on this subjectivity, we arranged a manual reinterpretation of every main log pass in fold 2 of our dataset, with the goal of comparing this reinterpretation with the official interpretation in our dataset. We gave this task to a well integrity researcher with a decade of experience in well logging engineering and research, who is not part of the team behind the official interpretations. He carried out his interpretations directly on the log data files without first having seen the official interpretation. To display and process the log data files in order to make his interpretations, he used the WellCAD software by Advanced Logic Technology. (Similar software like Techlog by Schlumberger and Geolog by Emerson E&P Software could also have been used for the same task.)

3.7. Baseline method setup

The random baseline described in Sec. 3.5 is not a very interesting lower baseline, as any working classifier can beat it. A more interesting baseline to compare the neural network results to would be a classifier using a very simple approach. For this baseline, we classified the BQ and HI parameters in each 1 m well segment i by a simple thresholding of the CBLF channel's median value $\widetilde{\mathrm{CBLF}}_i$ inside the segment. We chose CBLF as the input as it is a simple channel that provides good overall information on the well status.

For an interpretation parameter with K classes $C_0,\ldots,C_{K-1}$, we define an ordered sequence of thresholds $T_0, \ldots, T_K$, where $T_0 = 0\,\text{mV}$, $T_K \rightarrow \infty\,\text{mV}$, and $T_k <= T_{k+1}$. If $T_k <= \widetilde{\text{CBLF}}_i < T_{k+1}$, we assign class label C_k to segment i. To find these thresholds with a simple and virtually parameter-free method, we employed decision trees, specifically the implementation in the scikit-learn library by Pedregosa et al. (2011). We balanced the classes and specified the maximum number of leaf nodes as K. ($K = 6$ for BQ, and $K = 2$ for HI.)

We found the baseline results using 6-fold cross-validation as shown in Fig. 4(a). We held out one fold at a time for testing, and used the remaining folds to fit a decision tree. For each test fold, we compared the baseline interpretations with the official interpretations as described in Sec. 3.5. After finding results for each test fold in this way, we used the metrics' mean value as our overall result.

3.8. Neural network setup

All of our data channels are regularly sampled along a depth axis. Some data channels are 2D, being regularly sampled in time (the VDL channel) or azimuthal angle (the AWBK, IRBK, T2BK, AIBK, and UFLG channels) as well. For this type of data, convolutional neural networks have been found to be very effective, and our network setup is therefore based on these. Our networks are implemented in Keras (Chollet et al., 2015) with the TensorFlow backend (Abadi et al., 2015).

As we are using 13 different data channels of different types, it is natural to use multiple inputs in the neural network setup. However, using one convolutional branch for each channel would be very computationally expensive. For that reason, we divide the channels across three branches, shown in Fig. 5, according to their dimensionality, source tool, and depth resolution commonality.

The USIT branch receives the USIT image channels at a resolution of 3 in (7.62 cm) and 5°. We upsample channels originally sampled at 10° by nearest-neighbour interpolation, for the same reasons as described in Sec. 3.3.2. The branch also receives the USIT curves, which we array broadcast to the same shape as the image data. When provided to the network, the 10 different data channels are stacked like color channels in an image. The 1D branch receives the CBLF and GR channels at a resolution of 6 in (15.24 cm), and the VDL branch receives the VDL channel at a resolution of 4 in (10.16 cm) and 5 μs. Because the number of time samples in VDL channels varies across files, we trimmed all channels to 240 time samples (1.2 μs), the lowest common value. VDL channels with fewer samples (provided by uncommon tools) were discarded.

All data channels are normalised individually before input, to a mean value of 0 and a standard deviation of 1, to avoid channels with higher values being weighted more. Missing data channels and data channel values are replaced with zero-values. To augment the data, we exploit the periodic azimuthal symmetry of the USIT images by rolling and flipping the images in angle. When doing so, we also change the values of angular curve channels (AZEC, UCAZ, and RB) accordingly.

The network setup follows recommendations by Chollet (2018). Each branch contains convolutional layers and maxpooling layers, and we tuned their size and number based on our accuracy metrics. The convolution kernel sizes are 3 × 3 for the USIT branch, 7 for the 1D branch, and 5 × 5 for the VDL branch. The poolings of the maxpooling layers are 2 × 2 in the USIT branch, 2 in the 1D branch, and 2 × 4 in the VDL branch. As recommended by Chollet (2017, 2018), we used depthwise separable convolution layers. These use a representionally efficient convolution approach that separates spatial and channel convolution kernels, and gave us better results than conventional convolutional layers. The three branches are merged after global average pooling, whereupon densely connected layers are used for classification. The convolutional and dense layers use the ReLU activation function. To combat overfitting, the dense layers use a dropout of 0.5, whereas the convolutional layers use a spatial dropout of 0.2. We used the RMSprop optimiser with a learning rate of 0.001 and a training batch size of 16 samples. To balance the classes and because training often quickly reaches its highest accuracy (see Sec. 4.2), we defined an epoch as consisting of 3000 samples drawn equally from every class.

Conventionally, neural network classifiers for problems with K classes use softmax activation and categorical crossentropy loss with an output vector of length K. Element k' of the target vector Y is $Y_{k'\neq k} = 0$, $Y_{k'=k} = 1$, where C_k is the manually labelled class for the interval

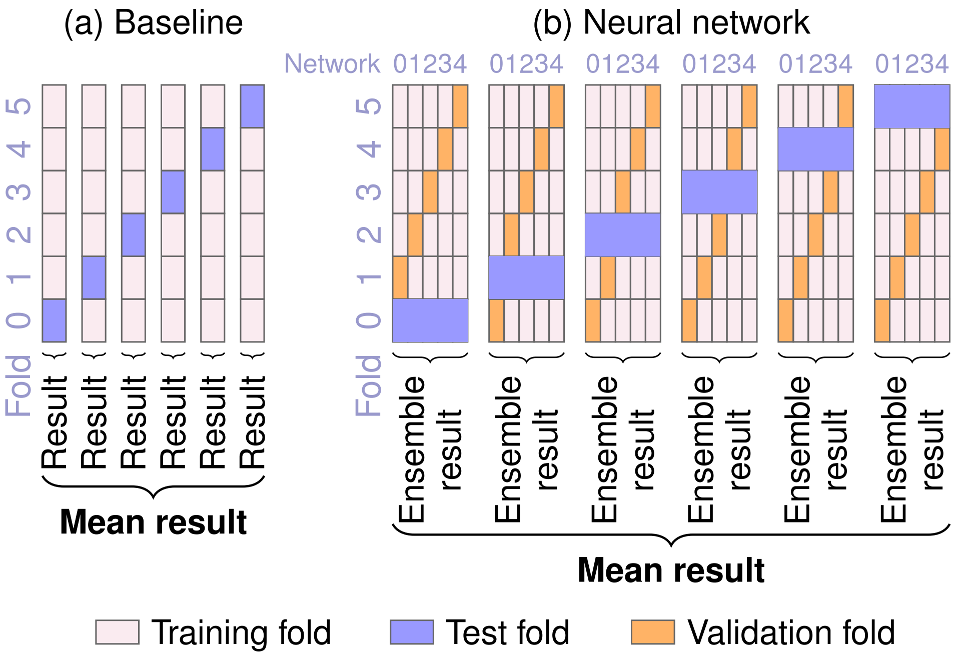


Fig. 4. Usage of folds for (a) the baseline case, which uses 6-fold cross-validation, and (b) the neural network case, which uses a form of ensembled cross-validation. For each test fold in the latter case, 5 networks were trained using different validation folds. Together, these networks form an ensemble that was tested on the test fold.

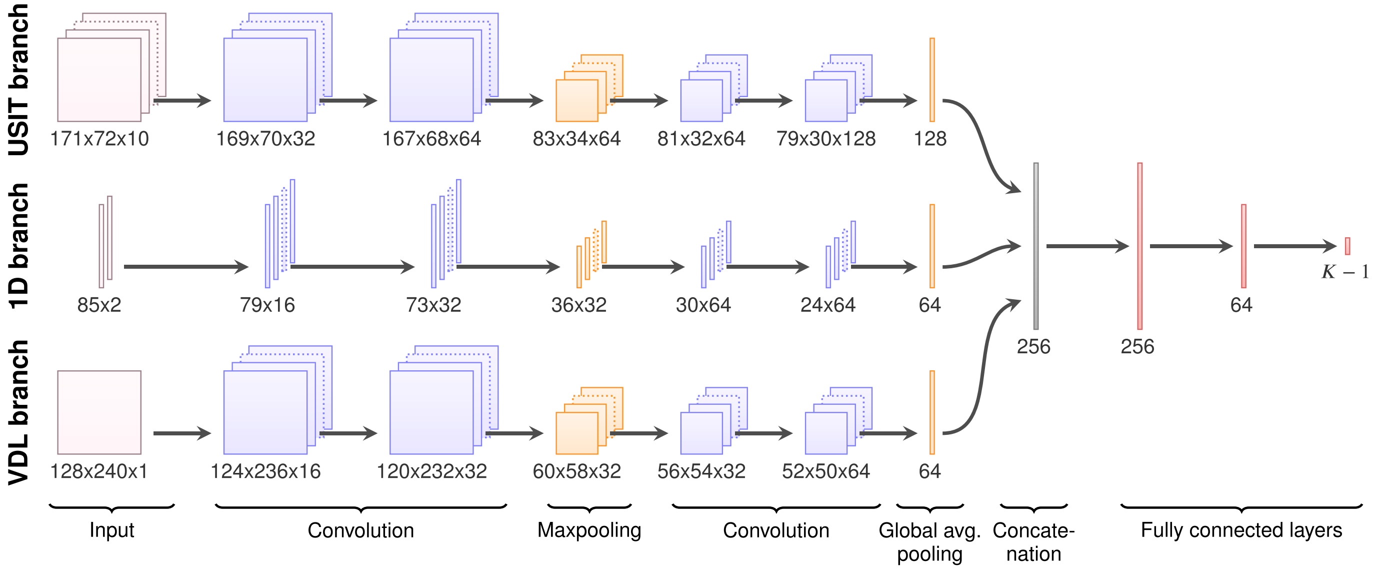


Fig. 5. Setup of the neural networks, from the three input branches to an output whose size is given by the number K of classes.

(Chollet, 2018). However, this approach implies that the classes are nominal and would ignore the ordinal nature of our BQ classes. Instead, we used an approach like that of Cheng et al. (2008), with a $(K-1)$-length target vector where $Y_{k'<k} = 1$ and $Y_{k'\geq k} = 0$, using sigmoidal activation and binary crossentropy loss. From the first k' in the output vector for which $Y_{k'} < 0.5$, we determine the predicted class as $C_{k'}$. This choice of target vector ensures that the loss function is lower the closer the class prediction is to the manually labelled class. We observed that this choice increased the BQ accuracy of our network.

For training and evaluation, we used a form of ensembled cross-validation, shown in Fig. 4(b). As we did for the baseline, we hold out one fold at a time for testing. Here, however, we also hold out one fold at a time for validation in order to select the best-performing network state during training, i.e., the epoch with the highest balanced precise accuracy on the validation fold. Thus, for each test fold, we use the five different validation folds to train five BQ and five HI networks. The five networks in each group are ensembled to combine their predictions. We used the median of the class predictions of the ensemble's component

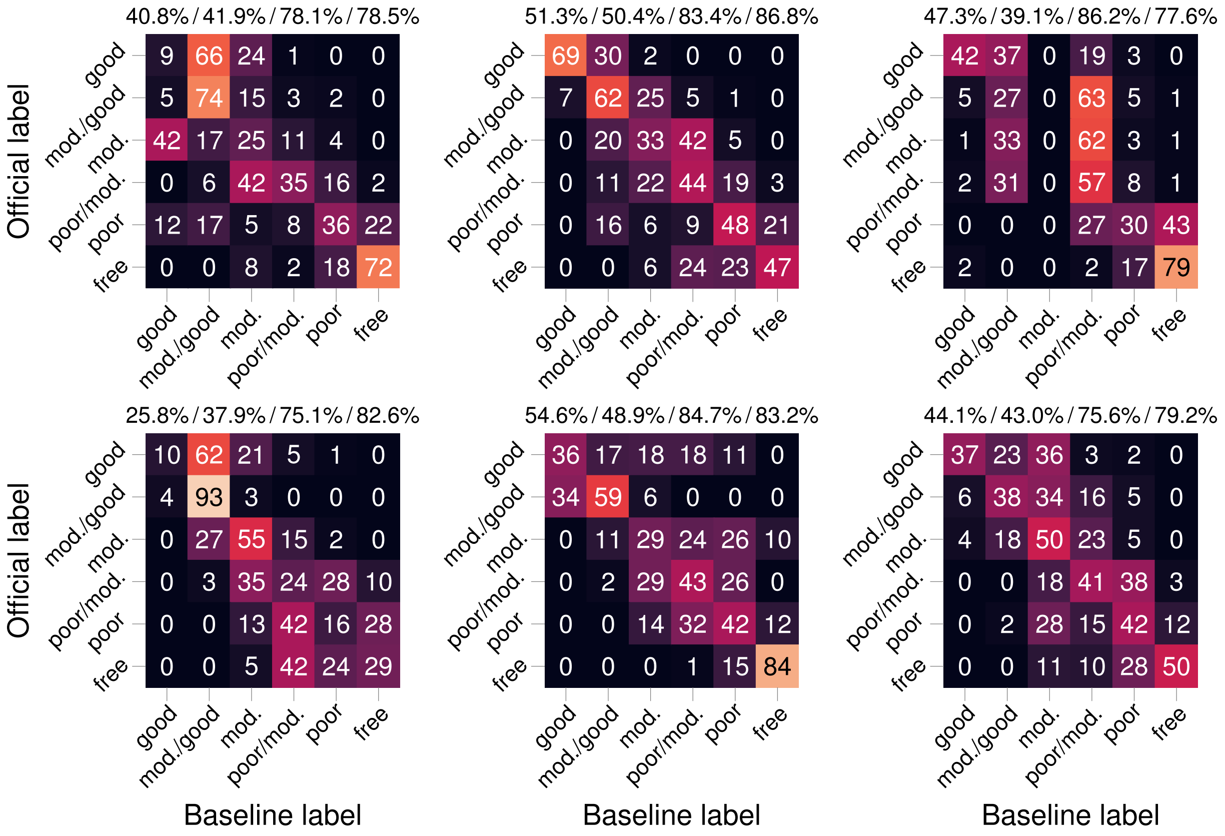


Fig. 6. Bond quality confusion matrices for the baseline method, showing the six test folds in reading order. Each matrix is normalised so that each row sums up to 100%. The numbers above each matrix represents its accuracy metrics UPA/BPA/UAA/BAA.

networks as a combination rule, as da Costa (2014) suggests. (With an odd number of networks in the ensemble, the median always unambiguously provides an average class. It is also equivalent to majority voting if there is an absolute majority in the ensemble.) We found the ensembles' accuracy metrics on their test folds and used their mean values as our overall result.

4. Results

We will first discuss the results of the baseline method, the neural network method, and the manual reinterpretation individually, before we compare them qualitatively for a specific well log in Sec. 4.4.

4.1. Results of baseline method

As Fig. 4 shows, we tested the baseline method on each fold individually, using the remaining folds for training. Figs. 6 and 7 show the BQ and HI results, respectively, for each test fold. Fig. 8 shows the overall BQ and HI results, found by averaging the results for the individual folds.

Table 3 shows the CBLF thresholds that the baseline method found to separate between different classes. When using fold 2 as the test fold and the rest for training, the BQ decision tree was unable to separate the 'Moderate' class from the 'Moderate to good' class. As we can see from Table 3 and Fig. 6, the baseline method could not and did not predict the label 'Moderate' for this test fold. (This minor issue could have been circumvented by increasing the maximum number of leaf nodes in the BQ decision trees from 6 to 8, but we found that this would have led to an overall slight decrease in accuracy.)

The overall results show that a simple thresholding of the median CBLF value within the segment can predict the official manual labels surprisingly well, predicting the same BQ class or a class adjacent to it 80.5% of the time, and the same HI class 81.2% of the time. This easily outperforms the corresponding accuracies for a random baseline, which are 44.4% and 50%. We can also see in Figs. 6 and 7 that there is a large variation between the folds. For example, the baseline method agrees very well with the manual 'Good' BQ labelling in fold 1, while there is little such agreement in folds 0 and 3.

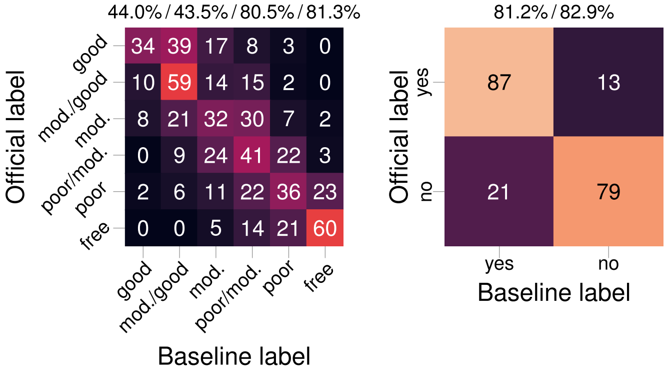


Fig. 8. Overall confusion matrices for the baseline method, found by averaging the results from all test folds, for bond quality (left) and hydraulic isolation (right).

Table 3
The CBLF thresholds in mV found by the baseline method to separate between the different classes for bond quality and hydraulic isolation.

Test fold	BQ					HI
	T_1	T_2	T_3	T_4	T_5	T_1
0	3.2	8.3	20.3	27.2	39.7	18.2
1	3.2	10.7	18.5	27.2	39.7	23.8
2	3.1	12.6	12.6	27.2	36.3	18.2
3	3.2	12.4	23.3	31.8	39.7	18.2
4	3.1	8.6	20.3	27.4	39.7	18.2
5	3.4	10.7	20.5	27.1	39.7	18.0
Mean*	3.2	10.6	20.6	28.0	39.1	19.1

*Excluding BQ T_3 for fold 2

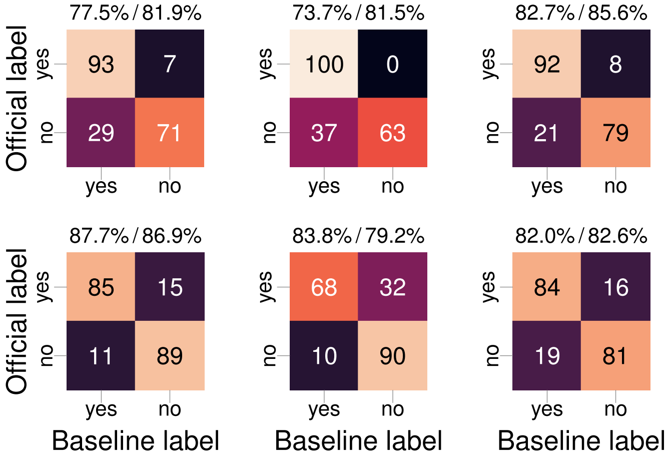


Fig. 7. Hydraulic isolation confusion matrices for the baseline method, showing the six test folds in reading order. Each matrix is normalised so that each row sums up to 100%. The numbers above each matrix represents its accuracy metrics UPA/BPA.

4.2. Results of neural network method

The training process for neural networks introduces stochasticity, through random initialisation and random choice of training samples. To investigate the significance of this stochasticity and to get a representative end result, we performed five repetitions of the training and testing process shown in Fig. 4(b). In other words, for every combination of test fold t and validation fold v shown in the figure, we trained five networks r, which gives us five ensembles r for every choice of test fold t. In the following, we present mean values and standard deviations based on these five repetitions.

Figs. 9 and 10 show the results for BQ and HI, respectively. We can see that there is still a great deal of variation between the folds, just as we saw for the baseline method. Fig. 11 shows the overall BQ and HI results, found by averaging the results for the individual folds.

Table 4 compares these overall results, including standard deviations over the five repetitions, with those of other methods. We can see that the neural networks perform significantly better than the baseline. We discuss possible factors that may still limit their performance in Sec. 5.

We found from the training histories that overfitting, where further training does not improve validation accuracy, happens quite early. After training, each network's state was chosen at the epoch where its validation accuracy peaked, which occurred after a median of 10 and 4.5 epochs for BQ and HI, respectively. Our definition of an epoch, explained in Sec. 3.8, means that even 10 epochs represents only 30000 samples, around half of the dataset.

We also found considerable variation in accuracy between networks trained, validated and tested with the same exact setup. To quantify this, we find the deviations between the accuracy metrics of individual networks and the metrics' average over all repetitions, and then find the root mean square of these deviations: If $a_m(v,t,r)$ is mth accuracy metric for repetition r of training a network using validation fold v and test fold t, and $\bar{a}_m(v,t) = \frac{1}{5}\sum_r a_m(v,t,r)$ is the mean over all 5 repetitions, then we

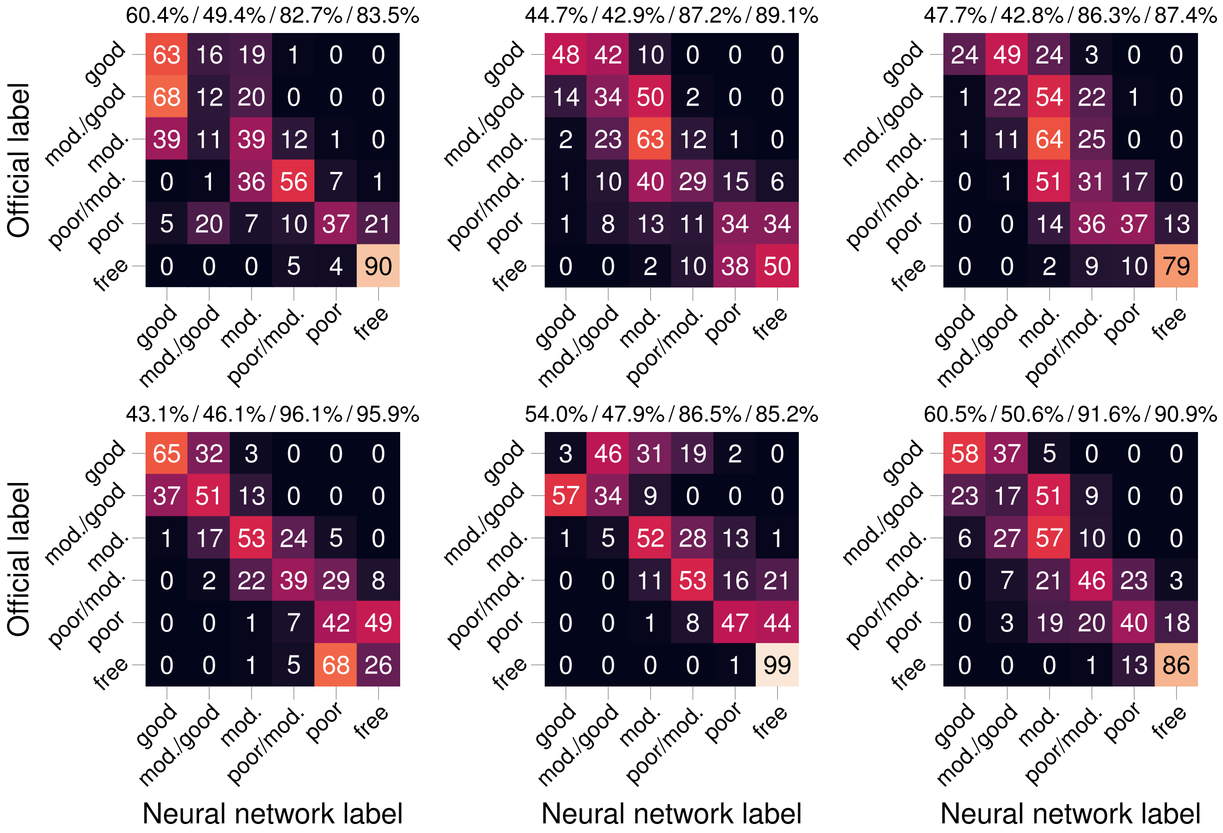


Fig. 9. Bond quality confusion matrices for the neural networks, plotted as in Fig. 6. Each confusion matrix is an average of the results of five repetitions.

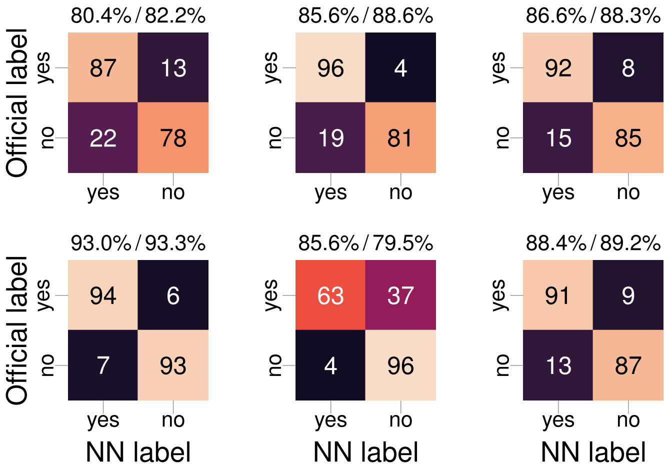


Fig. 10. Hydraulic isolation confusion matrices for the neural networks, plotted as in Fig. 7. Each confusion matrix is an average of the results of five repetitions.

calculate a network root mean square deviation as

$$\sigma_m^{\text{net}} = \sqrt{\frac{1}{6}\sum_t \frac{1}{5}\sum_{v\neq t} \frac{1}{5}\sum_r \left[a_m(v,t,r) - \bar{a}_m(v,t)\right]^2}\,. \tag{2}$$

The results for all accuracy metrics are shown in Fig. 12.

Similarly, we can look at the variation in accuracy between ensembles tested with the same folds. If $a_m(t,r)$ is the mth accuracy metric for repetition r of training an ensemble with test fold t and $\bar{a}_m(t) = \frac{1}{5}\sum_r a_m(t,r)$ is the mean over all repetitions, then we can calculate an ensemble root mean square deviation as

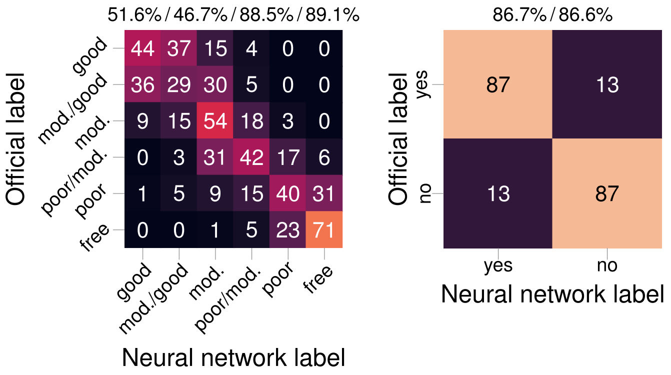


Fig. 11. Overall confusion matrices for the neural networks, found by averaging the results from all test folds, for bond quality (left) and hydraulic isolation (right).

Table 4

Comparison of the overall BQ and HI results of three methods: Expected values from random guessing (RND), the baseline method (BL), and neural networks (NN). We only report standard deviation for the neural networks, as it is the only method involving stochasticity.

Accuracy metric	BQ			HI		
	RND	BL	NN	RND	BL	NN
UPA [%]	16.7	44.0	51.6 ± 0.8	50	81.2	86.7 ± 0.3
BPA [%]	16.7	43.5	46.7 ± 0.7	50	82.9	86.6 ± 0.3
UAA [%]	44.4	80.5	88.5 ± 0.2	–	–	–
BAA [%]	44.4	81.3	89.1 ± 0.4	–	–	–

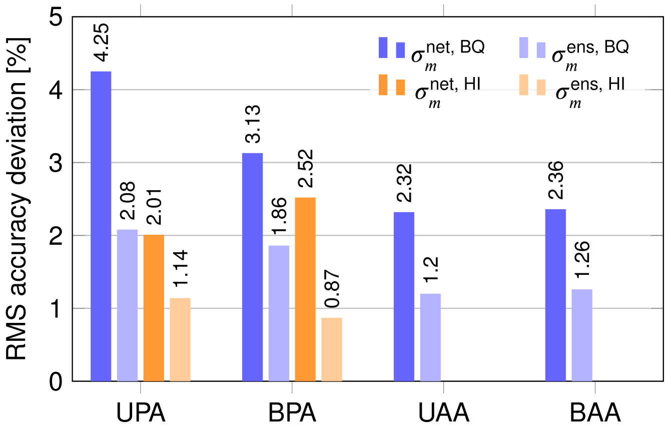


Fig. 12. Variation in accuracy metrics when training under the same conditions, shown as root mean square deviations for individual networks (eq. (2)) and ensembles (eq. (3)).

$$\sigma_m^{\text{ens}} = \sqrt{\frac{1}{6}\sum_t \frac{1}{5}\sum_r \left[a_m(t,r) - \bar{a}_m(t)\right]^2} \, . \qquad (3)$$

The results shown in Fig. 12 indicate that the ensembling roughly halves the variation, but even for the ensembles there is still quite a bit of variation left from repetition to repetition. We discuss this further in Sec. 5.1.

4.3. Results of manual reinterpretation

Fig. 13 shows the match between the official interpretation and the manual reinterpretation of each main log pass in fold 2. For the same data, it also shows the match between the official interpretation and the mean results of the five neural network ensembles trained for fold 2 as described in Sec. 4.2. (As this data represents a subset of fold 2, the latter results are similar but not identical to those shown for the same fold in Figs. 9 and 10.)

The results show that the reinterpreter tends to rate BQ and HI lower than the interpreters behind the official interpretation do, as the confusion matrices have more weight in their upper triangulars than in their lower triangulars. From the BQ confusion matrices, we see that the reinterpreter has a better agreement with the official interpretation than the neural network does on two out of the three most common classes ('Good' and 'Poor', but not 'Free pipe'), which also leads to a higher unbalanced precise accuracy. However, the reinterpreter's tendency towards significantly lower ratings leads to a lower adjacent accuracy than the neural network. The HI confusion matrix reinforces this picture. While the reinterpreter has a stricter interpretation of HI, disagreeing with 75% of the 'Yes' labels in the official interpretation, the neural networks almost always agree.

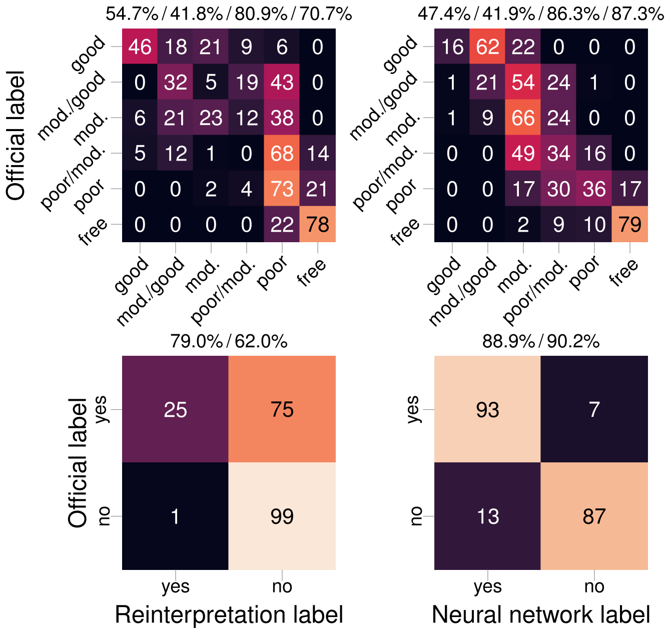


Fig. 13. BQ (upper) and HI (lower) confusion matrices comparing the official interpretation with the manual reinterpretation (left) and the neural network interpretation of the same data (right).

4.4. Qualitative comparison

The previous result sections present quantitative results. We will now take a more qualitative look at the interpretations for a specific well log. Fig. 14 shows the Volve 15/9-F-11 B well, which we also used as an example in Sec. 2.1.2. Here, the two rightmost columns compare BQ and HI interpretations, respectively. Note that this well log is a challenging case, with regions of both good and poor match between interpretations. The overall match between the official interpretation and the other interpretations is generally lower than the average over all well logs.

Looking at the official interpretation first, we see that it generally consists of long interpreted intervals interspersed with shorter intervals. These shorter intervals signify fluid patches in well bonded sections or short well bonded stretches in sections that otherwise contain channeling. An unusual feature in parts of the lower half of this log is that we see higher CBLF values at the same depths as impedance readings indicating well bonded solids around the entire cross-section. As explained in Sec. 2.1.1, we would expect to see low CBLF values together with such impedance readings. The official interpretation explains this as caused by micro-annuli to which the sonic tool can be overly sensitive, and it therefore trusts the ultrasonic tool over the sonic tool where they disagree.

Looking at the manual reinterpretation next, its quantitative match with the official interpretation in this well is a BQ UPA/UAA of 39.3%/ 58.4% (well below the overall reinterpretation results shown in Fig. 13) and a HI UPA of 74.5% (somewhat below the overall results in Fig. 13). Overall, the reinterpretation generally rates BQ lower than the official interpretation. While the HI match is very good from 3045 m and down, the match is worse in the 2911–3045 m section. Here, the official interpretation gave an isolating rating to several intervals that the reinterpreter rated as having 'Possible' HI, which, as Sec. 3.3.1 explains, is lumped into the 'No or uncertain' class. One particularly interesting disagreement is found around 2679–2735 m, which the official interpretation interprets as isolating solids and the reinterpretation interprets as non-isolating due to high azimuthal heterogeneity that may represent a fluid channel through the solids.

The neural network ensemble used here was the most average-performing of the five trained ensembles. Its interpretation's match with the official interpretation in this well can be quantified as a BQ UPA/UAA of 49.7%/73.1% and a HI UPA of 68.5%, all below overall results from Table 4. The BQ match is good down to 2910 m, below which the neural networks may be confused by the high CBLF values. The HI match is also generally good except for the 2756–2901 m section, where the neural networks fail by reporting a section with obvious channeling as hydraulically isolating.

Finally, the baseline interpretation's match with the official interpretation was a BQ UPA/UAA of 34.9%/80.5% and a HI UPA of 67.7%, all below the overall results from Table 4 except for BQ UAA. As this well is part of fold 2, we can see that the baseline interpretation is missing the 'Moderate' class as explained in Sec. 4.1. The baseline interpretation is markedly unstable, having regions where it flips rapidly back and forth between two adjacent classes when the CBLF value is hovering around one of the thresholds that separate classes. There are also some spikes in

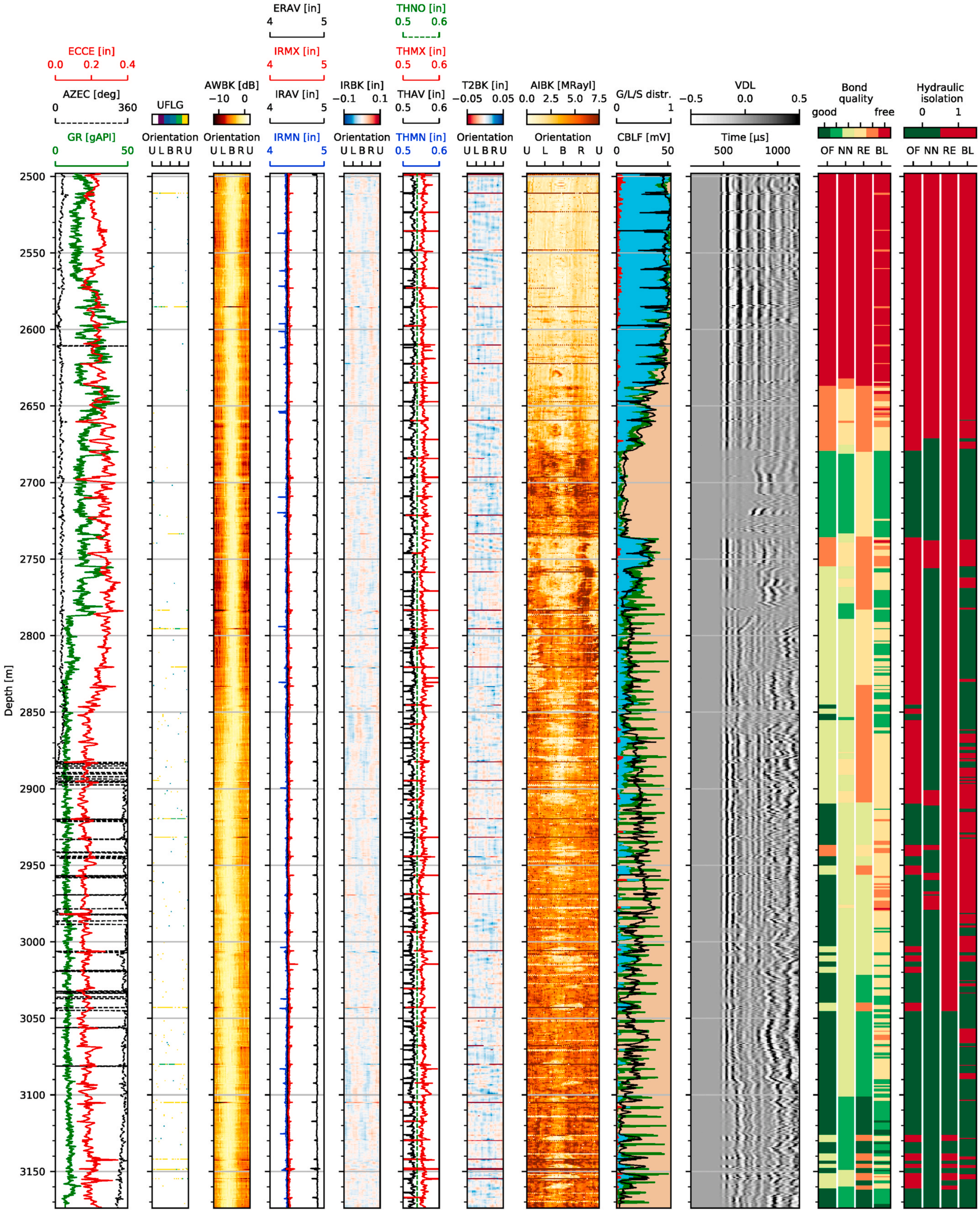


Fig. 14. Log plot of data from Volve 15/9-F-11 B, plotted as in Fig. 2, but with two extra columns comparing four interpretations of BQ and HI: The official interpretation (OF), the neural network interpretation (NN), the reinterpretation (RE), and the baseline interpretation (BL).

the interpretations related to the CBLF spikes caused by casing collars.

In general, both the neural networks and the baseline method give a good match with the official interpretation in the upper 3/5 of the log, except that both predict hydraulic isolation in a section with extensive channeling, which is a simple mistake to catch for a human correcting the automatic log. In the lower 2/5, however, both methods struggle to capture the official interpretation, probably due to the unusual combination of high CBLF values and good impedance readings.

5. Discussion

The accuracy metrics show that both the baseline and neural networks perform quite well, with the neural networks giving a 3–8% improvement over the baseline on every accuracy metric. It is somewhat surprising that the baseline still performs as well as it does while only being based on the median of the CBLF channel within each segment. However, as we discussed in Sec. 2.1.1, the CBLF channel, which is all the baseline method has access to, contains much of the same information as the acoustic impedance channel, which the neural networks also have access to. The results thus indicate that a simple thresholding of CBLF can be sufficient to make a decent interpretation, even though this can be improved significantly by also using information from other channels. It may also indicate that CBLF forms the backbone of the official interpretations in this dataset, which is consistent with what we know about the process behind these interpretations.

The manual reinterpretation gives an interesting perspective on interpretation bias. While there is no objective ground truth available to give us an objective reference for interpreters' biases, we can at least compare different interpreters' relative biases with each other. The confusion matrices in Fig. 13 indicate that the official interpreter team has a positive bias compared to our reinterpreter, or equivalently, that the reinterpreter has a negative bias compared to the official interpreter team. The neural networks, however, have been trained on the official interpretations and therefore tends to share the same bias. This may be the main reason that the neural networks outperform the reinterpreter when using the official interpretation as a reference. This also underscores that it is important for the training dataset to be thoroughly quality controlled interpreted log data, deliberately chosen to represent a reference for automatic interpretation.

Despite the promising results from the neural network, it is important to analyse which factors hold back its performance. In the following sections, we will discuss some factors that may limit the performance of our automatic interpretation system, and discuss other possible improvements.

5.1. Data heterogeneity

Consider a fairly homogeneous dataset, where the variation in the data is small compared to the size of the dataset. For such a dataset, we would expect to see only small variations between the results of different test folds. However, both the baseline method (Figs. 6 and 7) and the neural network methods (Figs. 9 and 10) show large variations in the results between different folds. This indicates that the dataset is strongly heterogeneous. In other words, the variation in our data seems to be large compared to the size of our dataset.

This is supported by other findings. The point of overfitting generally came quite early, which indicates that the relationships between data and labels that the networks learn from the training set have a limited generalisability to unseen data. Additionally, networks trained and tested with the exact same setup could end up with quite different accuracy on the validation and test sets, as noted in section 4.2. In other words, the random nature of network initialisation and training sample choice has a strong effect on how well the relationships learned by the networks can be generalised to unseen data.

We see several possible sources of this data heterogeneity, which we cover in the following subsections.

5.1.1. Data size

Perhaps the dataset is simply not large enough compared to the variation in the data? For example, some well or measurement conditions may be rare enough in our dataset that the networks are unable to discern meaningful relationships between data and labels from the available interpreted data.

The question is thus whether accuracy would increase with more data. While we have already used all the data we have available, we can turn the question on its head and investigate whether accuracy would *decrease* with *less* data. To do this, we reduced the size of the dataset by removing entire log operations from the folds while retaining the class balance as well as possible. The networks were then trained and validated on the reduced folds and tested on the full folds.

The overall BQ results shown in Fig. 15 for different reductions shows that even using as little as 38% of the dataset does not have a very strong impact on the accuracy. Thus, while having more data may reduce the variations between folds, the current performance does not seem to be primarily limited by the size of the dataset.

5.1.2. Subjective labelling in the dataset

In Sec. 2.2 we considered subjectivity in interpretation tasks, and more specifically inter- and intraobserver variability. The manual reinterpretation described in Secs. 3.6 and 4.3 and discussed at the beginning of the current section shows that the interobserver variability can be very strong in well log interpretation. We consider it likely that the official interpretations are also affected by interobserver (and possibly intraobserver) variability. In other words, our dataset may contain very similar input data samples that have been interpreted differently. In a case of two instances of similar data with different interpretations, a network trained on the first interpretation would have a difficult time predicting the second interpretation and vice versa. Additionally, a network trained on both would get mixed messages during training.

To investigate whether this effect limited our performance, we looked more closely at who performed the interpretations in our dataset. While the interpretations were to some degree produced as a team effort, the reports containing them also specify who the first interpreter was. He or she performed the initial interpretation, before it underwent quality control by other interpreters.

To investigate the effect of interobserver variability, under the assumption that this is not completely eliminated by the quality control process, we tried to reduce this variability by selecting a subset of our dataset with the most commonly used first interpreter. (This subset

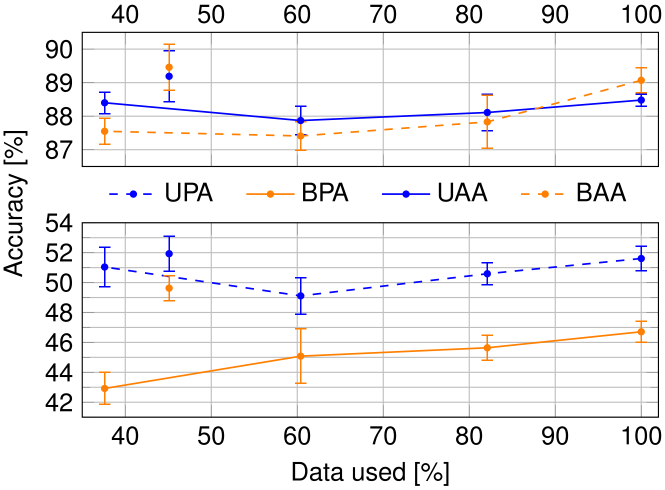


Fig. 15. Accuracy against the percentage of the dataset used, for the four different bond quality accuracy metrics. The points along the lines represent a simple reduction of the dataset, while the free points represent a subset of the dataset sharing the same first interpreter.

represents 45.1% of the total dataset.) We divided this subset into six folds and trained and tested neural network ensembles for BQ on the folds in the same way as before.

Fig. 15 shows significantly better results when all interpretations share the same first interpreter. We see particular improvement in the balanced accuracies, which weight the rarer intermediate categories 'Moderate to good', 'Moderate', and 'Poor to moderate' more than the unbalanced accuracies. We would expect these intermediate categories to be more subjective than the more clear-cut categories 'Good', 'Poor', and 'Free pipe'. Thus, our results indicate that the performance is limited by some degree of interobserver variability. To quantify interobserver and intraobserver variability further, however, a dedicated study would be necessary.

It may also be possible to reduce this subjectivity by using a different annotation system for the well log interpretations. For example, the system used for annotating bond quality uses a rating scale from 'Good' to 'Poor' (as well as 'Free pipe', which may at times be difficult to separate from 'Poor'), which is inherently opinion-based. An annotation system aiming for a more objective description of the distribution of material behind the casing may be able to result in more consistent annotations, although there may still be disagreements as to what that distribution is.

5.2. Other possible causes of reduced performance

Beyond the factors that may cause data heterogeneity, other factors may also have limited our performance.

5.2.1. Network setup

It could also be argued that the network setup is not ideal. However, we experimented with a number of variations, including reducing and increasing the capacity of the network. These changes most often did not affect the accuracy significantly, although some changes gave slightly negative effects. For example, while we could expect that choosing a lower learning rate would help reduce the accuracy variation shown in Fig. 12, our tests indicated that it did not, but instead reduced the overall accuracy slightly. This indicates that performance may mainly be limited by other factors than the network setup.

5.2.2. Differences in interval size

Manual interpretation often defines quite large depth intervals with the same interpretation. It can be argued that some of these intervals are coarse-grained and could be divided into multiple subintervals with different labels for BQ and HI. The two manual interpretations in Fig. 14 show this effect to some degree. The automatic interpretation, on the other hand, is in principle free to label each 1 m segment differently. In practice, both the baseline and the neural networks end up with intervals (by which we mean a series of consecutive segments with the same interpretation) that tend to be smaller than the manual interpretations. This discrepancy can reduce the match between the manual and automatic results.

To investigate the interval sizes, we looked at the length of interpreted intervals in the official manual interpretations used for training and testing, and the resulting neural network interpretations. Looking at BQ, the median official and neural network interval lengths are 29 m and 8 m, respectively. For HI, they are 26.5 m and 17 m. This shows that the neural networks do tend to perform a finer-grained interpretation of the well than human interpreters.

This limits our accuracy metrics, as they are calculated through segment-by-segment comparisons between the finer-grained automatic interpretations and the coarser-grained manual interpretations. To get an idea of how much these differences in interval size reduce the interpretation accuracy metrics, we tried forcing the neural network interpretations to use the same depth intervals as the reference interpretations. For all segments inside each official depth interval, we found the median of the automatic labels and set all labels to this median. We performed this procedure separately for BQ and HI.

Table 5 compares the original results shown Table 4 with results calculated from these interval-restricted automatic interpretations. We find that forcing our automatic interpretations into the coarser depth intervals used by the manual interpreters improves every accuracy metric by around 23%.

This result shows that the fine-grained nature of the automatic interpretations limits the reported accuracy when tested against coarse-grained manual interpretations. However, it also has implications for the training process, which is based on the same manual interpretations. If the manual interpretations are often too coarse-grained, this would make it more difficult for the networks to learn relationships between data and labels. As a hypothetical example, consider a coarse-grained manual interval with BQ labelled as 'Moderate' that also contains a smaller subinterval that would have been labelled 'Poor' if the manual interpreter had been requested to perform a finer-grained resolution. During training, the neural networks would be taught that segments similar to those inside this subinterval should be labelled 'Moderate' instead of 'Poor'. This would give the networks mixed messages when similar segments in other parts of the dataset are labelled 'Poor', likely reducing the networks' performance.

5.2.3. External information

The human interpreters may have access to information beyond what the data channels provide. For example, the well history can tell them where to expect the top of cement, as we saw in Sec. 2.1.2. The well log history may also tell them if some data channels should not be trusted uncritically, for example due to logs being run with an improper tool setup. The automatic system, on the other hand, only has access to the information present in the data channels. According to Benge (2014), such information can be quite important to the interpretation process.

However, there is a large variety of possibly useful external information, and it is is largely not present in machine-readable form in our dataset. It is therefore difficult for us quantify the importance of having such information available. In any case, however, a human interpreter is needed to verify an automatic log interpretation. Part of that task would be to compare it with any such information that might be available. Thus, the networks not having such information available would not be a major problem in this workflow.

5.3. Estimates of confidence

When using an automatic interpretation system like the one described in this article in practice, it would be useful if the system gave an estimate of its confidence in its interpretations. For example, an interval marked with low confidence could warrant closer scrutiny than an interval marked with high confidence.

However, while it is straightforward to get confidence estimates from conventional neural networks performing nominal classification, these confidence estimates are often not useful unless found using special techniques (Guo et al., 2017; DeVries and Taylor, 2018). Additionally, for the ordinal classification technique that we use to improve performance as described in Sec. 3.8, Beckham and Pal (2017) explain that estimating confidence is less straightforward. Additionally, the

Table 5
Comparison of the original neural network accuracy metrics and the metrics found when using the median neural network interpretation within each manual depth interval.

	BQ		HI	
Acc. metric	Original	Median	Original	Median
UPA [%]	51.6 ± 0.8	54.0 ± 1.6	86.7 ± 0.3	89.5 ± 0.5
BPA [%]	46.7 ± 0.7	50.0 ± 2.8	86.6 ± 0.3	88.6 ± 0.7
UAA [%]	88.5 ± 0.2	90.8 ± 0.4	–	–
BAA [%]	89.1 ± 0.4	91.3 ± 0.6	–	–

aforementioned special techniques are designed for nominal classification, and cannot be easily adapted to our approach.

Instead, we attempted to estimate the confidence of an ensemble through the agreement between the five individual networks that make up the ensemble. We tried quantify this agreement by a segment-by-segment agreement metric based on Cohen's weighted kappa extended to multiple raters (Cohen, 1968; Conger, 1980). However, for this agreement metric to be useful, it would need to be strongly negatively correlated with the prediction error, i.e. the distance between the predicted and reference labels, so that high agreement tends to coincide with low error and vice versa. Using Spearman's rank correlation coefficient ρ to quantify this cross-correlation, we found $\rho = -0.025$ for BQ and $\rho = -0.14$ for HI, i.e., a very weak negative correlation. Thus, we found that the disagreement between the individual networks in the ensemble unfortunately cannot be used to identify intervals requiring additional scrutiny. To identify such intervals, if it is even possible, another approach would have to be found.

6. Conclusion

Well log interpretation is a challenging problem, and so is creating an automatic well log interpretation system. In this work, we show how it is possible to train deep neural networks to interpret well logs through supervised learning: We show them a dataset of well log data and their interpretations and let the networks themselves draw the connections between data and interpretations. We cannot directly train the networks to come up with 'true' answers; there is no ground truth available, as the logs often do not unambiguously show what the true well status is. This is why we must use a dataset of manually interpreted data as the networks' reference for how well logs should be interpreted.

One particular limitation of such an approach is that the networks may not be able to interpret data that is not similar to data they have seen before. The neural networks can only make decent interpretations of edge cases, such as logs with strongly broken symmetry where the tool or casing is highly eccentered, if the training dataset contains enough examples of such data. Manual or automatic quality control of new logs to be interpreted might be necessary, as a neural network will give an output even for data of such poor quality that it should not be interpreted. Additionally, any supervised learning system would not be able to handle a type of data it has not seen before, such as data from a new logging tool. To learn to handle new types of data, the system would need to be trained on large amounts of interpreted data of the new type, although machine learning techniques such as transfer learning may reduce the amount needed.

Another major challenge is interpretation subjectivity. As the log data may be ambiguous, different interpreters may come to different conclusions based on similar data and apply different labels. This can give the neural networks mixed messages during training. For this reason, the dataset should ideally be thoroughly quality controlled data, hand-picked to form an internally consistent reference for training. Using a more objective annotation system instead of the inherently subjective rating scale used in our dataset may also help with this internal consistency.

Despite such challenges, the neural network interpretation results are very promising: For bond quality, we found an unbalanced precise accuracy of 51.6% and an unbalanced adjacent accuracy of 88.5%, and for hydraulic isolation, we found an unbalanced precise accuracy of 86.7%. Comparing the performance of the networks and a skilled interpreter on a subset of the dataset, we find that the networks' interpretations match the reference interpretations better than the manual reinterpretations do, according to five out of six accuracy metrics. While a comparison with a single reinterpretation is not sufficient to let us conclude that our networks generally agree better with the reference than manual reinterpretations do, it does indicate that the neural networks' overall ability to recreate unseen interpretations from the dataset is, at the very least, comparable with the ability of other skilled interpreters.

Credit author statement

Erlend Magnus Viggen: Conceptualisation, Methodology, Software, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualisation, Project administration. **Ioan Alexandru Merciu:** Conceptualisation, Investigation, Resources, Writing – original draft, Writing – review & editing, Funding acquisition. **Lasse Løvstakken:** Conceptualisation, Methodology, Writing – review&editing, Funding acquisition. **Svein-Erik Måsøy:** Conceptualisation, Writing – review & editing, Funding acquisition.

Data availability

Part of the dataset underlying this article is taken from the well integrity logs in the Volve Data Village dataset from Equinor (2018), hosted at https://data.equinor.com/.

Acknowledgements

We would like to thank Equinor for agreeing to release the data used in this study. We also appreciate input and feedback from Andreas Østvik, Erik Smistad and the rest of the CIUS machine learning team, and from Kevin Constable, Rune Godøy, and others on the Equinor team. Funding: This work was supported by the Research Council of Norway [grant number 237887].

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: large-scale machine learning on heterogeneous systems. URL: https://www.tensorflow.org/.

Akkurt, R., Conroy, T.T., Psaila, D., Paxton, A., Low, J., Spaans, P., 2018. Accelerating and enhancing petrophysical analysis with machine learning: a case study of an automated system for well log outlier detection and reconstruction. SPWLA 59th Annual Logging Symposium, p. 25. London, UK.

Albawi, A., De Andrade, J., Torsæter, M., Opedal, N., Stroisz, A., Vrålstad, T., 2014. Experimental set-up for testing cement sheath integrity in arctic wells. OTC Arctic Technology Conference. Offshore Technology Conference, Houston, Texas, p. 11. https://doi.org/10.4043/24587-MS.

Allouche, M., Guillot, D., Hayman, A.J., Butsch, R.J., Morris, C.W., 2006. Cement job evaluation. In: Nelson, E.B., Guillot, D. (Eds.), Well Cementing, second ed., pp. 549–612 Schlumberger. (chapter 15).

Anderson, W.L., Walker, T., 1961. Research predicts improved cement bond evaluations with acoustic logs. J. Petrol. Technol. 13, 1093–1097. https://doi.org/10.2118/196-PA.

Anthimopoulos, M., Christodoulidis, S., Ebner, L., Christe, A., Mougiakakou, S., 2016. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. IEEE Trans. Med. Imag. 35, 1207–1216. https://doi.org/10.1109/TMI.2016.2535865.

API, 1991. Recommended Digital Log Interchange Standard (DLIS), Version 1.00. Technical Report. American Petroleum Institute. http://w3.energistics.org/rp66/V1/Toc/main.html.

Beckham, C., Pal, C., 2017. Unimodal probability distributions for deep ordinal classification. Proceedings of Machine Learning Research 70.

Belozerov, B., Bukhanov, N., Egorov, D., Zakirov, A., Osmonalieva, O., Golitsyna, M., Reshytko, A., Semenikhin, A., Shindin, E., Lipets, V., 2018. Automatic well log analysis across Priobskoe field using machine learning methods. SPE Russian Petroleum Technology Conference, p. 21. https://doi.org/10.2118/191604-18RPTC-MS. Moscow, Russia.

Benge, G., 2014. Cement evaluation - a risky business. SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers, Amsterdam, p. 10. https://doi.org/10.2118/170712-MS.

Bennis, M., Torres-Verdín, C., 2019. Estimation of dynamic petrophysical properties from multiple well logs using machine learning and unsupervised rock classification. SPWLA 60th Annual Logging Symposium. The Woodlands, TX, USA, p. 11.

Bigoni, F., Pirrone, M., Pinelli, F., Trombin, G., Vinci, F., 2019. A multi-scale path for the characterization of heterogeneous karst carbonates: how log-to-seismic machine learning can optimize hydrocarbon production. SPWLA 60th Annual Logging Symposium. The Woodlands, TX, USA, p. 9.

Cheng, P.M., Malhi, H.S., 2017. Transfer learning with convolutional neural networks for classification of abdominal ultrasound images. J. Digit. Imag. 30, 234–243. https://doi.org/10.1007/s10278-016-9929-2.

Cheng, J., Wang, Z., Pollastri, G., 2008. A neural network approach to ordinal regression. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence, pp. 1279–1284. https://doi.org/10.1109/IJCNN.2008.4633963. IEEE, Hong Kong, China.

Chollet, F., 2017. Xception: deep learning with depthwise separable convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Honolulu, HI, pp. 1800–1807. https://doi.org/10.1109/CVPR.2017.195.

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

Chollet, F., et al., 2015. Keras. URL: https://keras.io.

Cohen, J., 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol. Bull. 70, 213–220. https://doi.org/10.1037/h0026256.

Conger, A.J., 1980. Integration and generalization of kappas for multiple raters. Psychol. Bull. 88, 322–328. https://doi.org/10.1037/0033-2909.88.2.322.

Crow, W., Williams, D.B., Carey, J.W., Celia, M., Gasda, S., 2009. Wellbore integrity analysis of a natural CO2 producer. Energy Procedia 1, 3561–3569. https://doi.org/10.1016/j.egypro.2009.02.150.

da Costa, J.D.P., 2014. Ensemble Methods in Ordinal Data Classification. Master's thesis. University of Porto.

Dai, B., Jones, C., Price, J., van Zuilekom, T., 2019. Auto-navigation of optimal formation pressure testing locations by machine learning methods. SPWLA 60th Annual Logging Symposium. The Woodlands, TX, USA, p. 10.

DeVries, T., Taylor, G.W., 2018. Learning Confidence for Out-Of-Distribution Detection in Neural Networks arXiv:1802.04865 [cs, stat].

Equinor, 2018. Volve data village dataset. Released under a CC BY-NC-SA 4.0 licence. https://data.equinor.com/.

Froelich, B., Dumont, A., Pittman, D., Seeman, B., 1982. Cement evaluation tool: a new approach to cement evaluation. J. Petrol. Technol. 34, 1835–1841. https://doi.org/10.2118/10207-PA.

Gkortsas, V.M., Bose, S., Zeroug, S., 2019. Machine learning for the automated detection of diagnosis-revealing features on leaky flexural wave imager data. 45th Annual Review of Progress in Quantitative Nondestructive Evaluation. American Institute of Physics, Vermont, USA, 050008. https://doi.org/10.1063/1.5099774.

Graham, W., Silva, C., Leimkuhler, J., de Kock, A., 1997. Cement evaluation and casing inspection with advanced ultrasonic scanning methods. SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers, San Antonio, Texas, p. 11. https://doi.org/10.2118/38651-MS.

Grosmangin, M., Kokesh, F.P., Majani, P., 1961. A sonic method for analyzing the quality of cementation of borehole casings. J. Petrol. Technol. 13, 165–171. https://doi.org/10.2118/1512-G-PA.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. Proceedings of Machine Learning Research 70.

Gupta, K.D., Vallega, V., Maniar, H., Marza, P., Xie, H., Ito, K., Abubakar, A., 2019. A deep-learning approach for borehole image interpretation. SPWLA 60th Annual Logging Symposium. The Woodlands, TX, USA, p. 10.

Hayman, A., Hutin, R., Wright, P., 1991. High-resolution cementation and corrosion imaging by ultrasound. SPWLA 32nd Annual Logging Symposium, p. 25.

Herold, B.H., Edwards, J.E., van Kuijk, R., Froelich, B., Marketz, F., Welling, R.W.F., Leuranguer, C., 2006. Evaluating expandable tubular zonal and swelling elastomer isolation using wireline ultrasonic measurements. IADC/SPE Asia Pacific Drilling Technology Conference and Exhibition, p. 11. https://doi.org/10.2118/103893-MS. Bangkok.

Hill, J.R., 1871. Improvement in Oil-Well Drilling. US Patent No, p. US112596A.

Jain, V., Wu, P.Y., Akkurt, R., Hodenfield, B., Jiang, T., Maehara, Y., Sharma, V., Abubakar, A., 2019. Class-based machine learning for next-generation wellbore data processing and interpretation. SPWLA 60th Annual Logging Symposium. The Woodlands, TX, USA, p. 17.

Kyi, K.K., Wang, A.G.J., 2015. Issues with cement bond and cement evaluation logs - case studies from offshore Malaysia. International Petroleum Technology Conference, International Petroleum Technology Conference. Doha, Qatar, p. 10. https://doi.org/10.2523/IPTC-18538-MS.

Li, H., Liu, G., Yang, S., Guo, Y., Huang, H., Dai, M., Tian, Y., 2019. Automated resistivity inversion and formation geometry determination in high-angle and horizontal wells using deep learning techniques. SPWLA 60th Annual Logging Symposium. The Woodlands, TX, USA, p. 11.

Liang, L., Le, T., Zimmermann, T., Zeroug, S., Heliot, D., 2019. A machine learning framework for automating well log depth matching. SPWLA 60th Annual Logging Symposium. The Woodlands, TX, USA, p. 9.

Mau, M., Edmundson, H., 2015. Groundbreakers: the Story of Oilfield Technology and the People Who Made it Happen. Fast-Print Publishing.

Miller, D., Stanke, F.E., 1999. Method of analyzing waveforms. US Patent No 5, 859,811.

Morris, C., Sabbagh, L., Wydrinski, R., Hupp, J.L., van Kuijk, R., Froelich, B., 2007. Application of enhanced ultrasonic measurements for cement and casing evaluation. SPE/IADC Drilling Conference, Amsterdam, p. 15. https://doi.org/10.2118/105648-MS.

Onalo, D., Adedigba, S., Khan, F., James, L.A., Butt, S., 2018. Data driven model for sonic well log prediction. J. Petrol. Sci. Eng. 170, 1022–1037. https://doi.org/10.1016/j.petrol.2018.06.072.

Oruganti, Y.D., Yuan, P., Inanc, F., Kadioglu, Y., Chace, D., 2019. Role of machine learning in building models for gas saturation prediction. SPWLA 60th Annual Logging Symposium. The Woodlands, TX, USA, p. 13.

Østvik, A., Smistad, E., Aase, S.A., Haugen, B.O., Løvstakken, L., 2019. Real-time standard view classification in transthoracic echocardiography using convolutional neural networks. Ultrasound Med. Biol. 45, 374–384. https://doi.org/10.1016/j.ultrasmedbio.2018.07.024.

Pardue, G.H., Morris, R.L., Gollwitzer, L.H., Moran, J.H., 1963. Cement bond log-A study of cement and casing variables. J. Petrol. Technol. 15, 545–555. https://doi.org/10.2118/453-PA.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Peyret, A.P., Ambía, J., Torres-Verdín, C., Strobel, J., 2019. Automatic interpretation of well logs with lithology-specific deep-learning methods. SPWLA 60th Annual Logging Symposium. The Woodlands, TX, USA, p. 20.

Popović, Z.B., Thomas, J.D., 2017. Assessing observer variability: a user's guide. Cardiovasc. Diagn. Ther. 7, 317–324. https://doi.org/10.21037/cdt.2017.03.12.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252. https://doi.org/10.1007/s11263-015-0816-y.

Shao, W., Chen, S., Eid, M., Hursan, G., 2019. Carbonate log interpretation models based on machine learning techniques. SPWLA 60th Annual Logging Symposium. The Woodlands, TX, USA, p. 13.

Sinha, B.K., Zeroug, S., 1999. Geophysical prospecting using sonics and ultrasonics. In: Webster, J.G. (Ed.), Wiley Encyclopedia of Electrical and Electronics Engineering. John Wiley & Sons, pp. 340–365.

Tubman, K.M., Cheng, C.H., Toksöz, M.N., 1984. Synthetic full waveform acoustic logs in cased boreholes. Geophysics 49, 1051–1059. https://doi.org/10.1190/1.1441720.

Tubman, K.M., Cheng, C.H., Cole, S.P., Toksöz, M.N., 1986. Synthetic full-waveform acoustic logs in cased boreholes, II—poorly bonded casing. Geophysics 51, 902–913. https://doi.org/10.1190/1.1442148.

van Kuijk, R., Zeroug, S., Froelich, B., Allouche, M., Bose, S., Miller, D., le Calvez, J.L., Schoepf, V., Pagnin, A., 2005. A novel ultrasonic cased-hole imager for enhanced cement evaluation. International Petroleum Technology Conference, p. 14. https://doi.org/10.2523/IPTC-10546-MS.

Vrålstad, T., Saasen, A., Fjær, E., Øia, T., Ytrehus, J.D., Khalifeh, M., 2019. Plug & abandonment of offshore wells: ensuring long-term well integrity and cost-efficiency. J. Petrol. Sci. Eng. 173, 478–491. https://doi.org/10.1016/j.petrol.2018.10.049.

Wang, H., Tao, G., Shang, X., 2016. Understanding acoustic methods for cement bond logging. J. Acoust. Soc. Am. 139, 2407–2416. https://doi.org/10.1121/1.4947511.

Wu, H.H., Pan, L., Ma, J., Dong, W., Fan, Y., Lozinsky, C., Bittar, M., 2019. Enhanced reservoir geosteering and geomapping from refined models of ultra-deep lwd resistivity inversions using machine-learning algorithms. SPWLA 60th Annual Logging Symposium. The Woodlands, TX, USA, p. 8.

Zemanek, J., Caldwell, R., Glenn, E., Holcomb, S., Norton, L., Straus, A., 1969. The borehole televiewer—a new logging concept for fracture location and other types of borehole inspection. J. Petrol. Technol. 21, 762–774. https://doi.org/10.2118/2402-PA.