Virkkunen, Iikka; Koskinen, Tuomas; Siljama, Oskar

Virtual round robin 2 – Phased array inspection of dissimilar metal welds

Published in:
Nuclear Engineering and Design

DOI:
10.1016/j.nucengdes.2023.112555

Published: 01/12/2023

Document Version
Publisher's PDF, also known as Version of record

Nuclear Engineering and Design 414 (2023) 112555

Contents lists available at ScienceDirect

Nuclear Engineering and Design

journal homepage: www.elsevier.com/locate/nucengdes

Virtual round robin 2 – Phased array inspection of dissimilar metal welds

Iikka Virkkunen [a,*], Tuomas Koskinen [a], Oskar Siljama [b]

[a] Department of Mechanical Engineering, School of Engineering, Aalto University, Finland
[b] Trueflaw Ltd., Finland

ARTICLE INFO

Keywords:
Round robin
Ultrasonic
NDE
Virtual flaw
Dissimilar metal weld

ABSTRACT

Round-robin exercises have traditionally been laborious to arrange in non-destructive testing (NDT). The exercises have involved manufacturing of costly big mock-ups and then distributing them around the world to facilitate testing by numerous laboratories. This has limited both the number of such round robins and their scope. Often the round robins have contained small number of flaws and the representativeness of these flaws has been limited. Nevertheless, the few round robins that have been completed have yielded significant additional understanding on the capability of the used NDT methods and procedures.

Recently, the increased use of automated inspections together with the development of virtual flaws (independently by Trueflaw and EPRI) has enabled a new type of round robin, where instead of moving samples around the world, the round robin is focused on the data analysis and only pre-acquired data files are distributed. In 2019–2020, first of a kind virtual round robin (VRR) was completed. The round-robin allowed for the first time to compare inspection performance from teams around the world with statistically significant number of flaws and with ultrasonic data representative for nuclear dissimilar metal weld inspection. The study resulted in important new insight into NDE reliability for nuclear applications.

However, as a first of a kind study, the first virtual round robin also contained some significant limitations. In particular, the data sets distributed were limited in order to limit the effort needed from each participating inspector. The reduced amount of data acquired was compensated by using pre-optimized data gathering, possible only with prior knowledge on the flaws present. While these choices were well justified for the first round-robin, they also made direct comparison of VRR results and real-life inspector performance problematic. In addition, the first VRR focused primarily on flaw detection and the data was insufficient for sizing.

To address these shortcomings of the first round robin, a second round robin was completed in 2021–2022. In this second round robin, more representative data was used for evaluation. In addition, increased emphasis on the hard-to-detect small flaws was put forward to get improved into detectability especially in the low end.

The more representative data required much more significant effort from the inspectors, which reduced the participation as compared to the first round robin. Furthermore, the emphasis on difficult-to-detect cracks may have further deterred participation, as the exercise may have been seen as too challenging. While the number of downloaded data sets (23) was similar to previous exercise, the number of returned sets was reduced to 5, compared to previous 18. Despite the smaller than expected participation, the results revealed several interesting features. The results displayed marked variation. Also, the false call rate was significantly reduced, as compared to the previous study. This could be attributed to the more rich data set, which allowed more comprehensive evaluation and exclusion of potential false calls.

The recent advances in machine learning (ML) for ultrasonics also introduced an interesting opportunity to compare machine learning results with the human inspectors. Developing an optimized machine learning model for the present data was outside the scope of this study. Instead, an independently developed model, if somewhat sub-optimal, was used. Thus, the results should not be taken as a measure of ML performance as such. Nevertheless, the comparison between human results and ML model are informative and illustrate the potential benefits of automated data evaluation.

* Corresponding author at: Advanced Manufacturing and Materials, Department of Mechanical Engineering, Aalto University, Puumiehenkuja 3, 02150 Espoo Finland.
E-mail address: Iikka.Virkkunen@aalto.fi (I. Virkkunen).

https://doi.org/10.1016/j.nucengdes.2023.112555
Received 3 April 2023; Received in revised form 13 August 2023; Accepted 15 August 2023
Available online 21 August 2023

1. Introduction

Non-destructive evaluation (NDE) is used in wide range of industries to provide information of the current state of the component to be inspected and, in particular, to find and characterize cracks or flaws that develop and compromise the structural integrity of the component during later use. The value of the inspections increases with increasing capacity to reliably detect small flaws and to reliably size them, as this allows more time and flexibility in mitigating the found flaws. At the same time, the detectability of the flaws decreases with decreasing flaw size and thus the methods are pushed to their limits. Furthermore, many inspections retain an element of human judgement, as the final evaluation of an indication is too complicated to be captured in deterministic set of rules. This potentially adds variability to the assessment.

Round robin exercises are used across many fields to study reliability and consistency of measurement. In NDE, these typically focus on flaw detection and characterization, and in the following round robin is used to refer to such studies. Notable previous NDE round robins include the PISC studies (Crutzen et al., 1998; Doctor et al., 1995; Lemaitre et al., 1996; Lemaitre et al., 1996; Lemaitre et al., 1996; Reale and Tognarelli, 1995; Reale et al., 1995) and the more recently the cast austenitic stainless steel (CASS) round robin (Jacob et al., 2018; Kull, 2018). Round robins are relatively rare in NDE due to the high cost preparing representative flawed test pieces and then passing them around among participants.

In previous work (Virkkunen et al., 2021b), this traditional hinderance was avoided with data-only "virtual round robin" and the use of virtual flaws (Svahn et al., 2018; Virkkunen et al., 2014; Virkkunen et al., 2016).

1.1. Virtual round robin

The advantages of a virtual round robin are many: the use of virtual flaws means, that small number of test blocks (even one) can be extended to provide sufficient space for any number of flaws, limited only by the amount of evaluation time required. Likewise, the limited number of flaws can be extended by copying the flaw signals in different locations and different backgrounds. In addition, the flaw signals can be manipulated, to give somewhat different features (most notably, amplitude). In particular, this enables a statistically representative number of flaws that can facilitate quantitative performance evaluation, such as a hit/miss probability of detection (POD) evaluation.

The first VRR clearly showed, that the virtual flaws and shared data files can be successfully used to study NDE reliability. However, as a first of a kind study, the first implementation necessarily contained some limitations that made it difficult to directly infer real-world performance from the results of the VRR. Some of these limitations were knowingly introduced as part of design trade-offs. For example, the amount of data was reduced to reduce the required inspector effort and to invite greater inspector participation. Others were discovered in analyzing the results. For example, even with the sufficient number of cracks, the results contained ill-conditioned results that limited the use of standard POD methodology.

1.2. Limitations of the first virtual round robin

The Virtual Round Robin necessarily operates with data files only. The data is scanned only once by one single team and not by every team as in a traditional round robin. Consequently, variation in data acquisition is omitted from study. The virtual flaw process could be used together with a traditional round-robin to include data acquisition differences while still obtaining POD results for each inspector. At the same time, this would forfeit the savings obtained from circulating data files instead of physical test pieces. Also, it would be unclear, whether the differences arise from differences in scanning or in evaluation. Thus, the virtual round robin and virtual flaws should not be considered as a mutually exclusive alternative to a traditional round robin. Rather, depending on the resources available and focus of interest the two can be effectively combined to study NDT performance.

Similar issues are found in the used procedure. In a traditional round robin, the teams can typically select their own procedure and scan data. For the present round robin, all the teams received the same data set and thus were constrained to use procedure associated with the available data. Again, alternative approach could be taken, where all the teams scan the data with their procedure and virtual flaws are still used to obtain sufficient flaw population for POD determination and this is a trade-off to be made considering the available resources and focus of interest.

Finally, for the first round robin, some compromises were made with the data, as explained above. The data was reduced to minimum viable to limit the effort required to participate and to ease data handling. To compensate for the limited data, the scan was made with prior information about the existing flaws. In both ways, the data differed from procedures actually used for such inspections. Thus, the POD values obtained could not be considered to be directly applicable to any specific inspection procedure. At the same time, it should be noted that procedures used for dissimilar metal weld inspections tend to be fairly similar and thus the effect of procedure differences can be expected to be minor.

1.3. Improvements for the second virtual round robin

While some of the identified limitations of the first VRR stem from the exclusive use of data, many others can be attributed to trade-offs made in the first VRR and can be improved on in further work.

Thus, the second VRR aimed to address these limitations. In particular, it was assumed that with the good results and positive experience from the first virtual round robin, it is now easier for the inspectors to justify participation in a follow-up, even if the inspector effort required is somewhat greater. Thus, some of the trade-offs of the first study would no longer be necessary. In addition, we can now take advantage of some of the lessons learned in the first VRR.

The most significant improvements relate to the data and flaws used. In the first VRR, simplified data sets were used, that contained only single channel of ultrasonic data, while normal field data is acquired with multiple refraction angles and/or scan lines to form a rich and somewhat overlapping data that allows detailed analysis. In this second VRR, the reduced data was replaced by full multi-channel data acquired according to industry standard acquisition procedures.

The source flaws for the first VRR comprised from a limited set of flaws used in qualification. This is problematic for the VRR as the qualification flaws are, in general, designed to be detectable and for POD evaluation also undetectable defects would be needed. This limitation was alleviated by introducing synthetic changes to the flaws, most importantly by reducing the signal amplitude. While these modifications did indeed result in reduced detection rates, as has been previously shown, (Koskinen et al., 2018), the very limited supply of raw flaw data made the files somewhat repetitive and may have influenced the inspector responses. This was further exacerbated by the very limited amount of raw flaw canvas data, that also needed to be recycled thus introducing further repetitiveness to the data.

Consequently, for the second round robin the data needed to be improved in multiple ways: more rich data sets, more raw flaws to start with, more difficult to detect small flaws and more raw data to be used as canvas. All this requires not only significant additional effort from the participating inspectors, but also vastly increases the effort needed to prepare such a data set. Furthermore, the high data quality was considered a key success factor for the second VRR and thus additional data was sought and existing data re-scanned multiple times until very high-quality data was obtained, even if this did cause some delays in preparation of the data files.

1.4. Probability of detection

One of the key advantages of using virtual flaws in this context is the ability to introduce sufficient number of flaws to test data and to gather statistically significant data that allow evaluation of inspection reliability. While the first VRR introduced high number of flaws, during evaluation it was found that the flaw distribution was still sub-optimal and in some cases precluded the use of standard POD methodology. The main cause of this was the limited supply of raw flaws, lack of small flaws and high performance of some of the inspectors. These issues are not limited to the VRR and similar issues are widely experienced in practical POD determination (Virkkunen, 2021).

In case of the first VRR, these limitations necessitated the use of non-standard techniques. As it turns out, there's several established methodologies on addressing such limitations. When the first VRR data was independently evaluated by several independent laboratories, the results diverged due to the use of different methodologies (Meyer et al., 2021). These discrepancies underlined the need for wider range of raw cracks and inclusion of small cracks for the second VRR.

1.5. Program for investigation of NDE by international collaboration (PIONIC)

Started in 2017, the Program for Investigation of NDE by International Collaboration (PIONIC) concentrates on the difficult inspection of dissimilar metal welds and related topics relating to inspection reliability. The project allowed the acquisition of suitable samples and also provided the international support needed for conducting a successful round-robin. The round robin is open for everyone, and industry participation is invited.

2. Materials and methods

As before, the virtual round robin was open to the industry.

2.1. The inspection target and procedure

The inspection case for the round robin was selected on the basis of the primary interest of the PIONIC project, and availability of suitable test blocks. On the second VRR, more emphasis was put on having sufficient and varied flaws and backgrounds to minimize repetitive patterns that might be recognized by the inspectors. On these grounds, data collected from VVER steam collector head provided by Fortum was used. While the weld details in VVER reactors differ slightly from typical western designs, the UT-data was considered sufficiently representative for present purposes. The weld geometry is shown in Fig. 1.

The flaw data, in contrast, was obtained from thermal fatigue cracks grown in dissimilar metal weld samples and from qualification flaw data.

While the procedures used by individual inspection companies for such inspections may differ, there is now a fairly well established "best practice" procedure for inspecting such welds. This practice has roots in the initial work by EPRI (Becker et al., 2004) and has sense been adopted by generic procedures (e.g. (Marois, 2011). In generic terms, these welds are inspected using phased-array longitudinal wave, about 1–2 MHz frequency and focusing to enable sufficient detection capability.

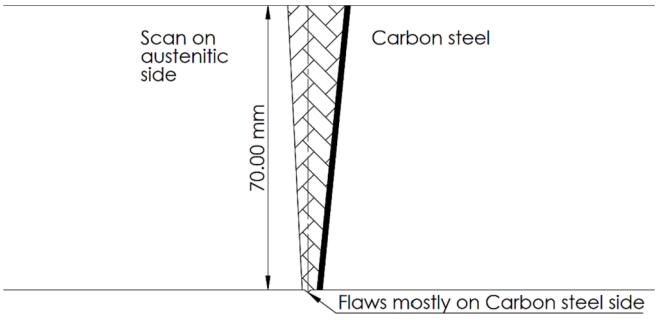

Fig. 1. The inspection target and the weld geometry.

For the purpose of the virtual round robin, we wanted to have a generic procedure that participants could choose to follow in their evaluation. The participating teams were free to deviate from this procedure or to use alternate (possibly internal) procedure for the evaluation. However, the data to be evaluated was already collected according to the generic procedure given and thus any differences in the procedure were necessarily limited to the evaluation of the existing data.

In the first VRR, number of participants provided insufficient information on the procedure they used and it was difficult to assess the sources of the differences in the evaluation. Thus, for this second VRR, the provided procedure is more detailed to allow a more consistent basis for evaluation. Also, the inspectors are explicitly required to complete the inspection using this procedure and are allowed to do differing evaluation in addition to the standard one. Thus, we hope to get consistent and comparable results as well as indication of potential improvements and/or alternate evaluations.

In contrast to the first VRR, the current data set provides much richer data representative of actual field inspection and sufficient for sizing procedures. While this was an important development, it also significantly increased the effort needed to develop the virtual data.

2.2. Mock-up and flaw population

The raw canvas data was acquired form a dissimilar weld mock-up provided by Fortum with identification "YD016".

Majority of the used flaws were manufactured to additional mock-ups and these flaws were added to the flaw population. Plate mock-ups were manufactured for the express purpose of providing additional flaw population for this project and provided by EPRI, designated as W2289 and W2660. Thermal fatigue cracks were manufactured to these mock-ups by Trueflaw. While scanning these additional mock-ups, it became evident that the thermal fatigue flaws produced by Trueflaw exhibit significantly lower amplitude than the previously used solidification cracks for the same nominal size. In addition, some flaws from qualification data were used.

Initially, it was planned to complete destructive examination on the flaw mock-ups to reveal the true depth of the cracks. However, in view of the current results, this was postponed. Furthermore, normally the flaw manufacturer provides dependable depth information based on destructive examination of validation cracks. Due the plan to destructively examine these cracks, this too was omitted. Consequently, the depth information available from the flaws used is somewhat uncertain. The depth information was estimated by the manufacturer based on similar cracks and should be considered indicative only. Table 1. lists the estimated flaw sizes for each sourced flaw.

The flaws were scanned with multiple tightly spaced scan lines (see below) to provide multiple different manifestations from each flaw. The

Table 1
Sourced flaws and estimated sizes.

Flaw id	Length (mm)	Depth (mm)
A	5.3 ± 0.5	1.8 ± 1
B	8.4 ± 0.5	2.8 ± 1
C	7.6 ± 0.5	2.5 ± 1
D	15.1 ± 0.5	5 ± 3
E	1.7 ± 0.5	0.6 ± 1
F	7.9 ± 0.5	2.6 ± 1
G	14 ± 0.5	4.7 ± 2
H	20 ± 0.5	6.7 ± 3
I	9 ± 0.5	3 ± 2
J	1 ± 0.5	0.3 ± 1

canvas data, in contrast, was much more sparce and thus each physical flaw provided multiple different (if related) views of the same flaw that could be used independently for virtual flaw implantation.

2.3. Data acquisition

The data was acquired through mechanized inspection with Dynaray Lite ultrasonic instrument. Phased array Transmit Receive Longitudinal (TRL) setup was utilized with 1.5 MHz 32 element matrix probes. Based on previous experiences from the first VRR, the setup was optimized for best possible data quality. Thus, the whole aperture of the probe was used to create refraction angles from 40° to 70° with a one degree step. The matrix orientation of the elements allowed implementation of the skew angle of −15° to + 15° with 3° step. The focal point of the ultrasonic wave was set at the bottom of the sample. The phased array law setup is demonstrated in Fig. 2. For canvas scanning, the refraction angle step needed to be decreased to 5° and −12° to 12° skew angles due to manual scanning. The recording frequency would be otherwise too low for quality data from manual probe movement.

As the whole aperture was used for increased accuracy, no electronic scanning could be used. Therefore, the probe was moved along the index direction mechanically or manually. The samples were scanned in three different ways. For flaw samples in plates, 20 index steps with 2 mm step size was used. For the mock-up sample the amount of index steps was five. Due to size restrictions the mock-up sample needed to be scanned in three different sections. The scanning setup for mock-up sample can be seen in Fig. 3. For canvas data, the index step was 5 mm with total of 3 index steps. The canvas data was scanned in four different sections due to sample size. The probe was moved manually along a plastic rail show in Fig. 4. After the scan, plastic rail was adjusted 5 mm and a new scan line was recorded. For all the scans continuous feed water system was applied to ensure proper coupling. Due to applying multiple index steps during scanning amount of data per one sample was approximately 10 Gb.

Example scan image from one of the plate samples can be seen in Fig. 5. The canvas mock-up also contained some pre-existing flaws. However, these proved ill-suited to be used for virtual flaw process due to various adverse conditions. While the weld geometry was also slightly different, the canvas sample was much noisier than the plate samples as can be seen in Fig. 6.

2.4. Virtual data generation and obfuscation

The eFlaw generation and creation of the data files for the virtual round robin were contributed by Trueflaw. The flaw data was extracted from the multi-channel data using the most recent generation of Trueflaw's eFlaw technology. This latest generation makes use of Trueflaw's recent machine learning developments and offers much more detailed flaw extraction and re-introduction with further reduced artifacts.

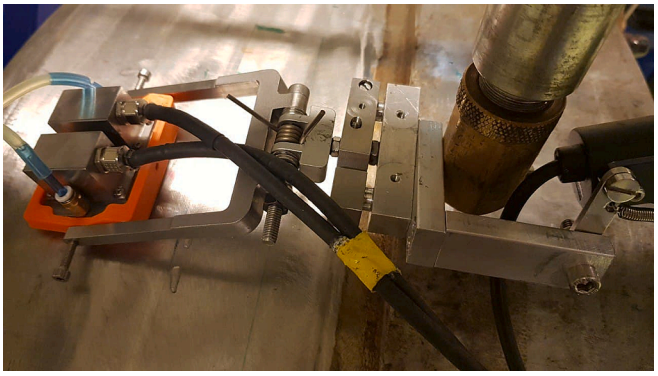

Fig. 3. The scan setup for the dissimilar metal weld mock-up.

Fig. 4. The cut off section for the canvas data with the plastic rail.

In contrast to the previous VRR, the current data included increased flaw free canvas data. This, removed the repetitive patterns and data modification artifacts that were left in the flaw-free canvas on the first VRR. Combined with the new generation eFlaw process, the data quality for this second VRR is considerably higher than for the first round robin, albeit with much more effort needed to complete the data generation.

The flaw source mock-up was much thinner than the final canvas mock-up and thus the signal amplitudes and signal to noise rates in the source mock-ups are expected to be higher than what would have been

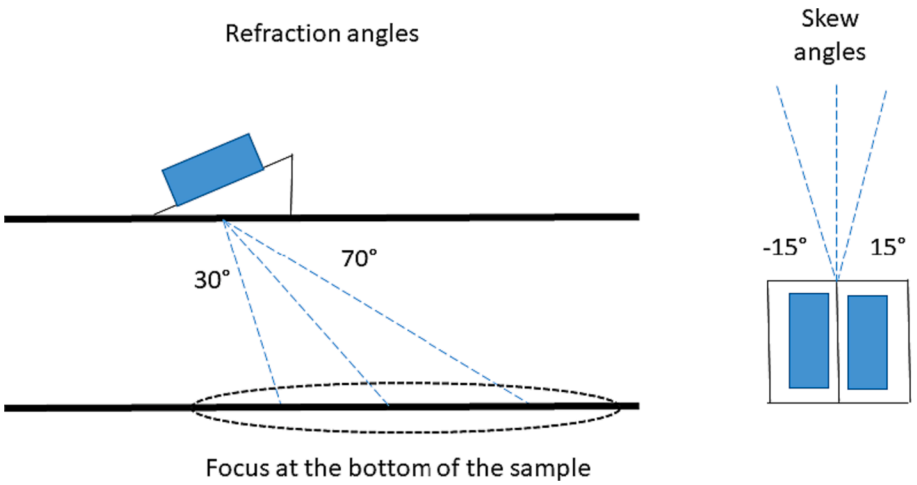

Fig. 2. Phased array setup.

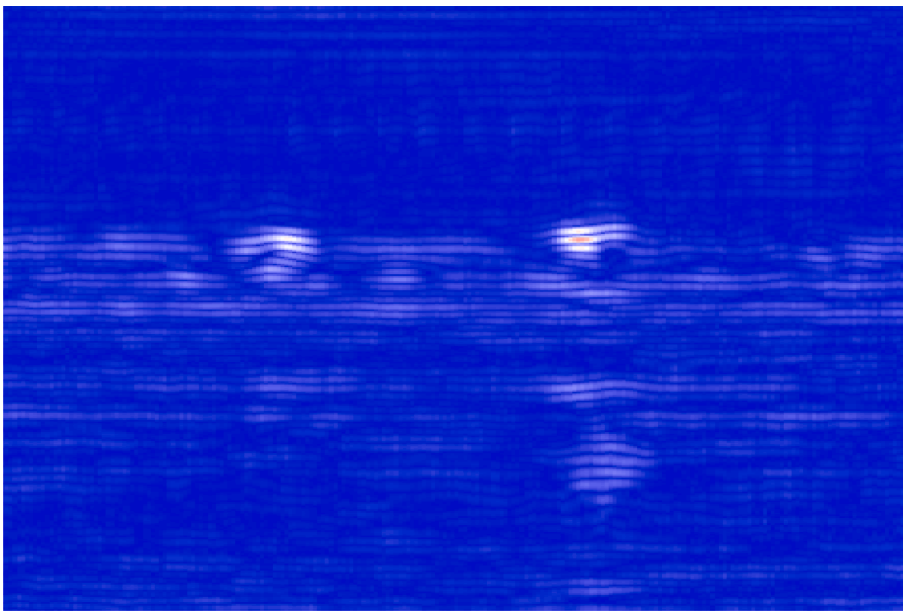

Fig. 5. Flaw specimens from plate sample W2289.

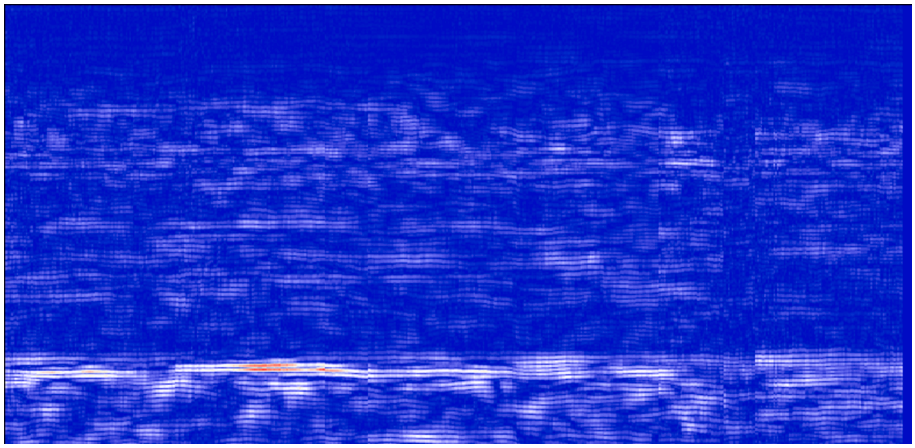

Fig. 6. Canvas scan data.

seen if similar flaws were scanned from the canvas mock-up. During the virtual flaw implantation, the amplitudes were varied to further increase the variability of flaw signals present in the mock-up. The variation was predominantly by factor below 1.0, i.e. the amplitude was decreased to provide more realistic targets.

Previous studies (Virkkunen et al., 2021b) have shown that variation in amplitude is an effective mean to introduce variation to detection tasks. However, for sizing the effect is less clear. Consequently, for sizing the results are reported both in terms of uncorrected size (i.e. the estimated size given in Table 1. for the source flaw) and in terms of corrected size, where the nominal size is multiplied by the applied amplitude factor.

In addition to the flaws sourced from actual mock-up, one simulated flaw was included. The flaw was simulated as 10 × 5 mm notch using CIVA ultrasonic simulation software. The simulated flaw (unaltered) was also provided in a separate file as reference flaw for the inspectors.

2.5. Evaluation of the data set

Before acceptance, a preliminary data set will be subjected to inspection by one of the authors, Tuomas Koskinen. As before, the trial run is to be completed as a blind exercise, that is Koskinen will have no prior knowledge of the location or number of flaws in the data files. Koskinen also did the scanning and thus was already familiar with the data. However, this time there's much larger canvas and flawed data variation and thus memorization effects can be expected to be minimal. The purpose of this trial evaluation is to confirm that the set-up works as expected and to identify any possible issues with the data.

2.6. Data delivery and collection of results

For data delivery, a web-app was used, as in (Virkkunen et al., 2021b). The web-app allows participants to register using their e-mail address and download the virtual round robin data files. The App randomly orders the files to make comparison of results more difficult and packs them in a single ZIP-compressed package. The package also contains a pre-filled Excel sheet for returning the results. The Excel sheets were designed to be machine readable to facilitate easy data collection and they automatically included identifier that tied the excel sheet to the data set it described. E-mail was used for registration to allow contacting the participant for delivering the results and for possible additional information, if needed.

In contrast to the previous round robin, the guidance now required the inspectors to provide evaluation using provided procedure and allowed additional evaluation using alternate procedure. None of the inspectors opted to submit multiple results.

The sizing requested the inspectors to detail the method used for sizing with the procedure giving two principal methods as follows:

"flaw tip" method: identifying the echo attributable to the highest flaw tip and determining its height using the analysis software.

"flaw image" method: estimated upper extent of the flaw image.

The flaw tip method should be preferred, when possible and flaw image used as a fallback method.

2.7. ML evaluation

While the round robin for human inspectors was the first and foremost aim for the VRR2, the round robin data also provided unique opportunity to be used for other studies. With the recent development of machine learning data analysis tools (Posilović et al., 2022; Cantero-Chinchilla et al., 2022; Bevan and Croxford, 2022; Tyystjärvi et al., 2022; Taheri et al., 2022; Cantero-Chinchilla et al., 2022; Shipway et al., 2021; Koskinen et al., 2021; Ye and Toyama, 2021; Siljama et al., 2021; Gantala and Balasubramaniam, 2021; Virkkunen et al., 2021a), it was also of interest to potentially use the data to see how ML would compare with human results and if ML would show potential to address some of the issues uncovered.

The used ML models are best developed for the specific procedure and data used in the inspection, but such model was not available for the present data.

EPRI and Trueflaw have previously collaborated to develop an ML model for dissimilar metal weld (DMW) inspection (to be published). For this study, the EPRI/Trueflaw DMW model was used. The model is designed for slightly different data, with multiple closely aligned index lines. To allow the usage of the model, the current data was repeated as if several index lines with the same data would have been acquired. This may increase the false call rate of the model, as random noise is repeated in consistent pattern and may be mistaken for a flaw. The virtual flaws used in the VRR data were also sourced partly from the same physical flaws that were used in development of the EPRI/Trueflaw DMW model. Thus the data separation is imperfect. However, leakage effects are expected to be minimal, as the virtual flaws were scanned with different set-up and the data is visually quite different.

The model evaluates each channel separately and the results are combined to for the final detection results.

3. Results

Results on the evaluation of the data set by one of the authors presented an opportunity to go conduct a post-analysis with the inspector and to gain deeper insight into some of the unexpected features of the. This proved valuable as these features were also observed in the submitted independent sets and thus these results are also presented in this section.

Altogether five responses were received for the virtual round robin. Out of those, three respondents submitted both detection and sizing results and two detection results only. The number of respondents is lower than expected and due to the small number cannot be expected to for a statistically meaningful sample of inspectors overall. Furthermore, the sample may be biased: as the number of downloaded data sets is much bigger than returned results, many potential participants decided not to participate only after they had had access to the data. They may have judged the data to be too challenging or time consuming to participate and so more confident, and thus presumably experienced, inspectors may be over represented in the received set.

The small number of submitted results and especially the ratio of downloaded data sets to the submitted result sets indicates, that the earlier concern about representative data imposing excessive burden on the participating inspectors are warranted. Indeed, there seems to be a trade-off, in particular for the most challenging inspections like the dissimilar metal weld, with obtaining realistic results and high participation.

Even if the sample is too small to estimate inspector performance in general, evaluation of the small set is highly illustrative. Despite the small sample, the results show great variation in some performance aspects. Small sample is expected to under estimate variation and thus the high variation is of interest in an industry with tight procedures in place aimed to minimize variation in inspection results. At the same time, the results reveal certain issues common to all or most of the result sets. Even with the small set, such commonalities are expected to have wider significance. Some of these issues have not been previously observed or published and have only been revealed now due to the unique setting made possible by the virtual round robin.

3.1. Data set evaluation detection results

The evaluation detection results are presented in Figs. 7 and 8. 23/62 flaws are successfully detected with 2 false calls. The number of misses is higher than would be expected and also the detection does not seem to correlate with the assumed crack size. Each of the missed cracks were located in the data and presented to the inspector and in each case, he indicated that the indication is clear flaw indication and should have been found. The missed cracks were typically such that the indications were only present on limited number of channels and especially on "non-central" channels with extreme skew or refraction angles. The inspector looked at both "merged" data and individual channels. Due to the variation in the material noise, merging channels to a single view obscured some flaws while not others. Thus, completing the inspection on merged data alone would yield unsuccessful result. It now appears, that the inspector tended, inadvertently, to focus more on the central channel and thus was liable to miss cracks with salient indications (when pointed out).

After the initial evaluation, some modifications were made to the file sets. Most notably, one included file (marked "Disqualified" in Fig. 7.) has large nominal crack size, but did not give corresponding indication. The flaw was a solidification crack sourced from an additional mock-up and due to these issues was removed from further study. In addition, some missed flaws considered overly challenging after evaluation and were removed or replaced. After the modifications, all the indications in the files were definitely discernable when pointed out.

3.2. Round robin detection results

The combined detection results are shown in Fig. 9. The detection results as a function of flaw sizes are shown in Fig. 10.

The detection results vary significantly, with hit rates ranging from 16 to 52 out of the total 65 defects in the data. The overall false call rate was lower than in the VRR1, with false call counts ranging from 3 to 18. This can be attributed to the more rich data that allowed the inspectors to more accurately exclude potential false calls from the data. There's no significant correlation between false call rate and detection performance, as seen in Fig. 11.

The data contained two simulated defect responses. These were found by all the inspectors. As in previous studies (Koskinen et al., 2021), the simulated defects appear to be significantly easier to detect than real defects, though the difference is likely pronounced in difficult microstructures such as these.

The detection results were also plotted as a function of nominal (and corrected) flaw sizes and traditional logistic POD curve fitted. The data does not show clear POD dependence on flaw size (either corrected or uncorrected). The results were also investigated in terms of indication amplitude, and this did not improve the dependence. Thus, in the present flaw size range, the detection is chiefly determined by factors other than flaw size or amplitude. These may include the noisiness around the

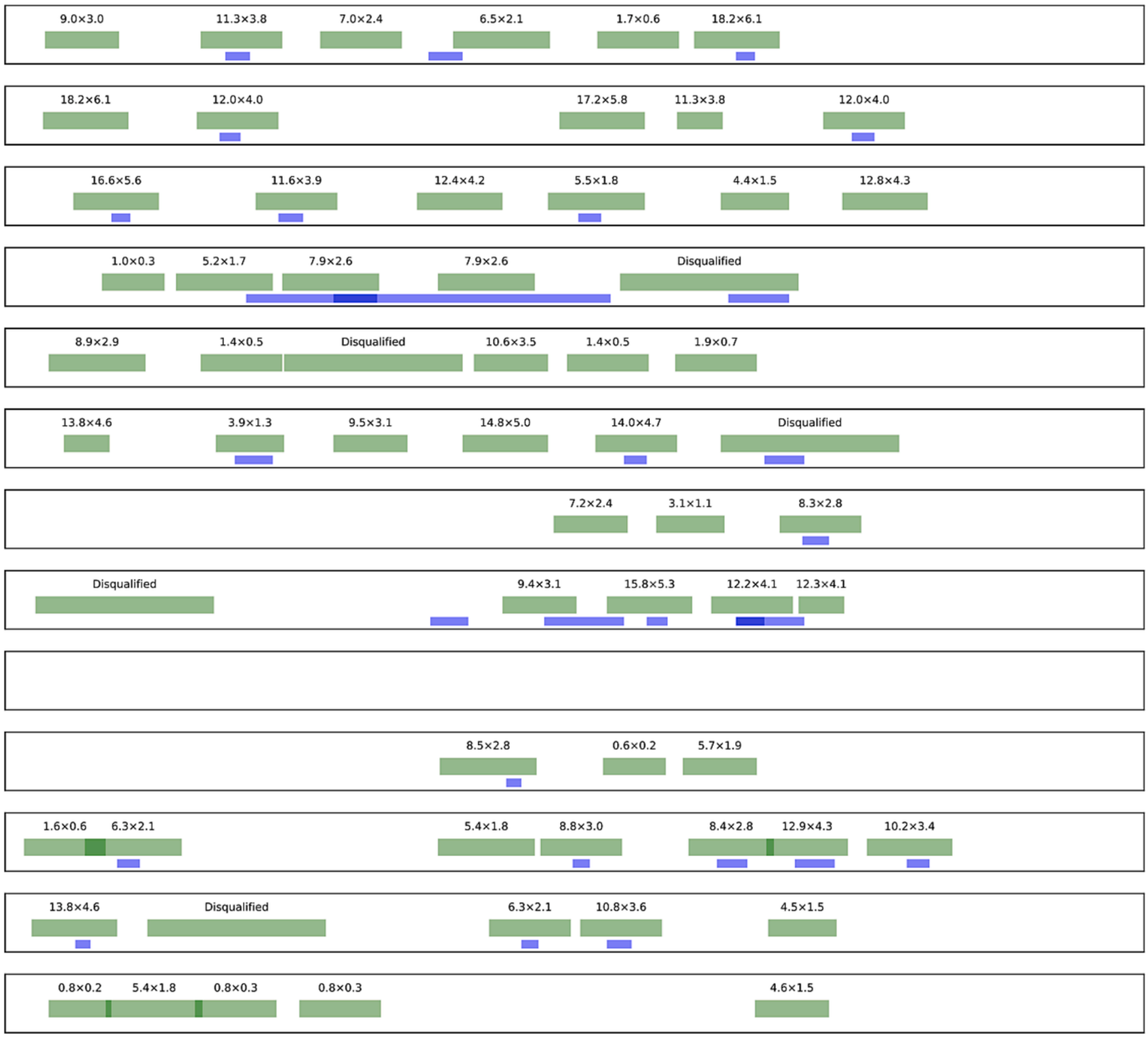

Fig. 7. Data set evaluation detection results. The flaws in data files are shown with thick green bars and inspector indications with thinner blue bars below. The flaws are annotated with nominal size (size of the source flaw corrected with amplitude factor), length × depth, in mm.

flaw location or interaction with the microstructure.

3.3. Round robin sizing results

The depth sizing results for the three submitted result sets are shown in Fig. 12. The sizing results are shown with the nominal size of the inserted flaw and with size corrected proportional to the amplitude factor applied.

Despite the uncertainties in the true depth information, it's evident that the sizing results do not correlate with the true depth in any of the result sets. While the results tend to be in a plausible range, neither the uncorrected or amplitude factor adjusted values show any correspondence to the depth estimates. The same physical flaws resulted in wildly different depth estimates when evaluated from different source scan lines and embedded in different locations. The differences cannot be explained by amplitude variation. They can only be attributed to uncertainty in the flaw sizing capability.

3.4. Machine learning detection results

The detection results obtained with the Trueflaw/EPRI DMW model are shown in Fig. 13. The model evaluation takes a couple of seconds to run on each data file. The POD results are not shown due to small number of misses.

While the ML results show higher false call rate than the human inspectors (23), the false call rate is still low enough for use as an aide for human inspector and not too far from the human inspectors. Also, the false call rate can to some extent be attributed to the model not being optimal for the present inspection. Furthermore, 4 of the 23 false calls coincided with false calls made by at least one other inspector.

The detection results, on the other hand, are significantly better than any of the human inspectors. The ML model detected 64/65 cracks. The single miss was also missed by all but one human inspectors and even there, the mark was near the end of the indication and may be a lucky false call. The missed flaw was not one close to the smallest nominal sizes. This indicates, that also the surrounding noise level has significant

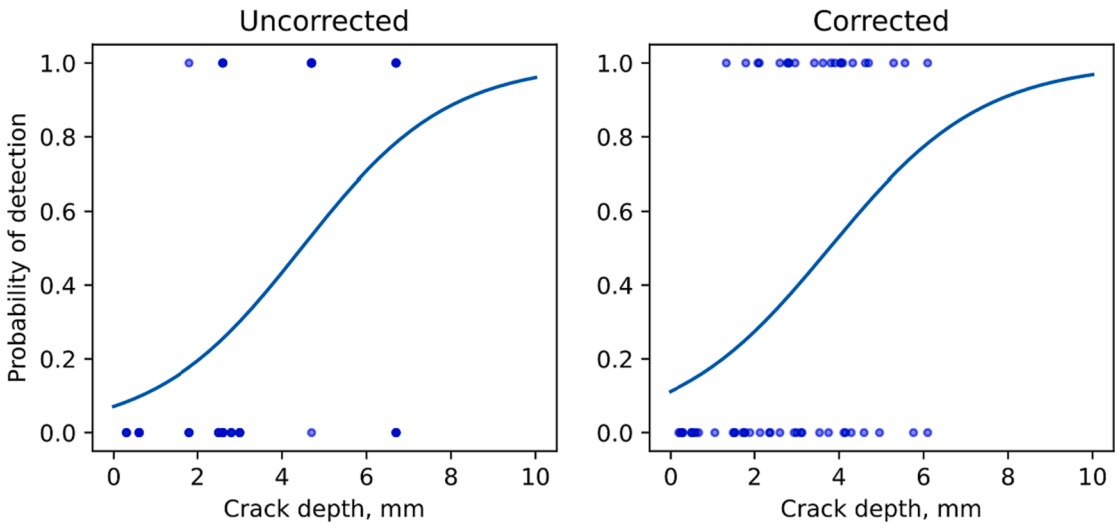

Fig. 8. Data set evaluation detection results as function of effective flaw size. In uncorrected plot (left) the flaw sizes are plotted as the nominal size of the source flaw. In the corrected plot (right) the flaw sizes are corrected with the amplitude factor applied. While the POD results display the expected rising trend, this is mainly due to consistent misses at the low end; even the biggest flaw sizes in the set are sometimes missed. Thus, no meaningful confidence bounds or $a_{90/95}$ can be obtained.

effect on flaw detectability.

4. Discussion

The ultrasonic inspection of dissimilar metal welds is notoriously challenging. The complex, coarse and generally unknown microstructural features of the specific weld under inspection may cause unexpected beam steering and spreading. This may cause variable damping effects, and also affect other indication properties. E.g. the strongest flaw signal may result from unexpected beam angle due to the combined effect of flaw properties and microstructural effect. The flaws themselves also grow through the microstructure and are expected to be affected by it. For example, a larger grain size, which is associated with more difficult UT inspection is also associated with higher SCC crack growth rates (Liu et al., 2021; López et al., 2006). In general, such compounding effects are difficult to anticipate and thus it is important to design the inspections to be robust against microstructural effects to the extent possible. To this end, the phased array procedures typically include high number of refraction angles and skew angles to maximize different opportunities for flaw detection. Often, also electronic scanning is used, but this was omitted in this study due to previous data showing that this would decrease flaw detectability in comparison to using the full probe aperture for beam forming and focusing.

The approach of using multiple channels and rich data afforded by the modern phased array equipment to maximize flaw detection and sizing performance is sound. At the same time, there's an important trade-off. With more data, the data analysis becomes more time consuming and tedious. Often, there's some time pressure on completing the data analysis, but even in the absence of explicit time pressure, the human attention is a scarce resource can be depleted (Temple et al., 2000) in tasks like inspection data analysis. The inspectors may use widely available technical measures, such as merged data, to alleviate the data overload, but the fairly simple tools commonly used are problematic for noisy inspections like the DMW. In particular, merging data can hide flaws that only appear on few channels and thus nullify the benefits of multitude of channels. Even when these are not used, the inspector may unintentionally focus more on the channels assumed more important. Thus, it appears that while adding more channels and data is well justified, it sees diminishing returns due to the excessive burden in data analysis. While these considerations are intuitively unsurprising, they have not, to our knowledge, been previously reported in this context and would have been difficult to study without the use of virtual flaws.

The virtual round robin data faithfully reproduces the general difficulties in DMW inspection. As the raw flaw data was scanned from high number of different scan lines and several different natural flaws, the variation in the weld microstructure caused the same flaw to give slightly different responses and to maximize in different refraction and skew angles. Thus, when the extracted virtual flaws were introduced to different locations in the canvas data with more limited number of scan lines, these correspond to more typical inspection data with some flaws appearing on more favorable position and others less so. Furthermore, as the flaws are embedded in different backgrounds and so variations in the flaw location in different noise conditions are represented.

It should be noted, that the flaws included in this study were focused on the difficult to detect size range and so even if the detection rate over this population is fairly low in some result sets, it should not be taken as an indication of poor overall performance.

Despite the emphasis on small difficult flaws, the best result sets show impressive performance with as much as 53/65 flaws correctly detected. There's significant variation in the results, as was also seen in the previous study (Virkkunen et al., 2021b; Meyer et al., 2021). The overall false call rate performance was very good, which can be attributed to the rich data available. As is often the case, there's no correlation between false call rate and detection performance.

Despite the differences in detection performance, all the result sets show decreased detection performance for flaws that are only visible in small number of channels and, in particular, "peripheral" channels with high skew angle. In addition, the local variations in surrounding noise level induce further variability to flaw detectability. Consequently, none of the flaw sizes present are reliably detected and the POD(a) curve does not display the expected rising trend over these data. The same effect was noted on the data evaluation by one of the authors, which allowed us to perform data review and gain further insight into the issue. The bigger missed cracks in the appearing in peripheral channels were clearly identifiable to the inspector, when pointed out. Thus, the present results indicate that with multichannel data like this, the inspectors are unable to lend sufficient focus on all the data and thus adding further data and channels will not provide the improved performance expected.

The ML results are not prone to this problem. The automated evaluation can easily be formulated in a way to allow consistent focus on all available data, as was done here, and the computational performance of modern hardware and ML models is good enough to allow comprehensive analysis. While the ML results provided in this study were not

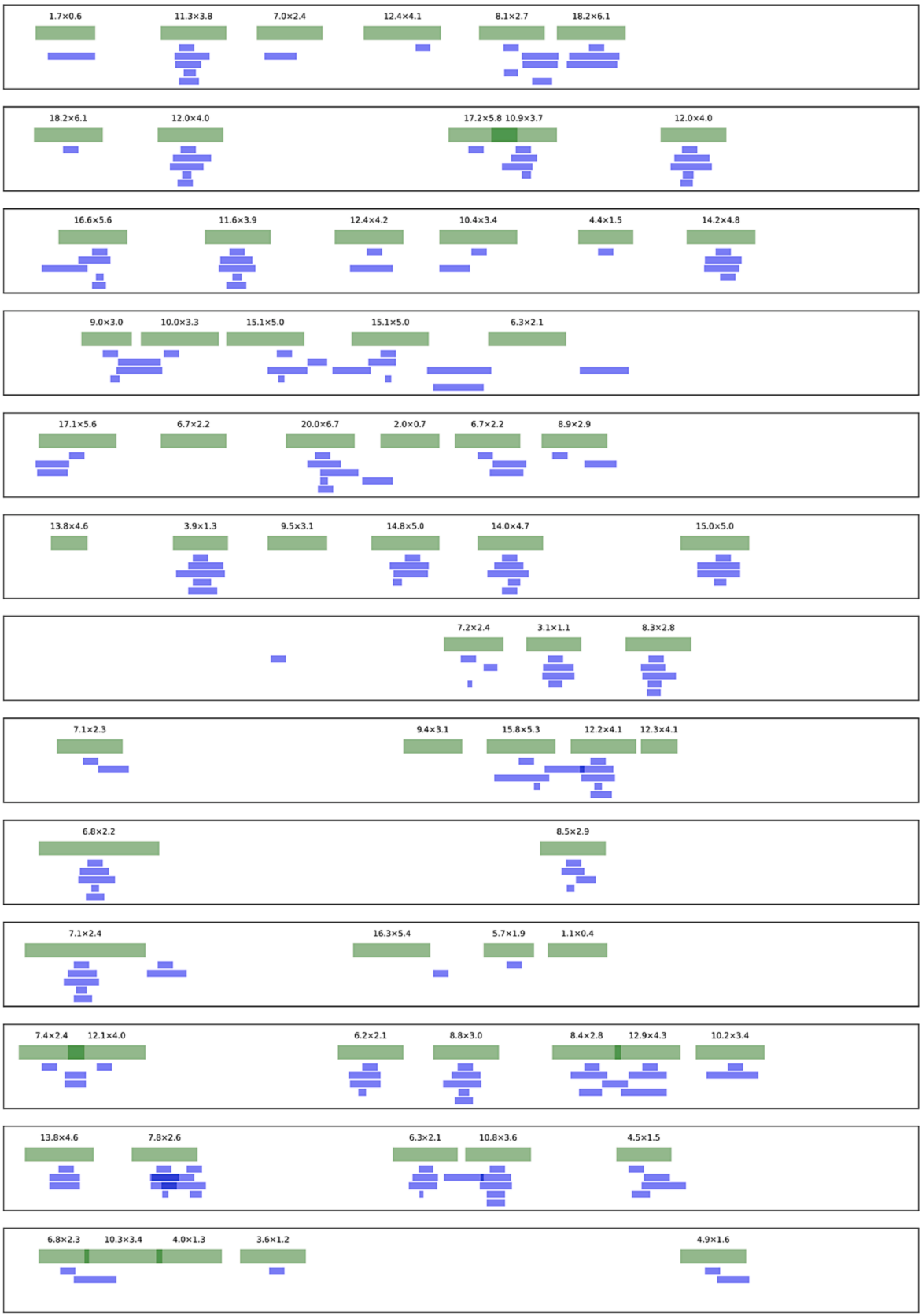

Fig. 9. Data set evaluation detection results. The flaws in data files are shown with thick green bars and each inspectors indications with thinner blue bars below. The results are ordered top to bottom best to worst in terms of hit count. The flaws are annotated with nominal size (size of the source flaw corrected with amplitude factor), length × depth, in mm.

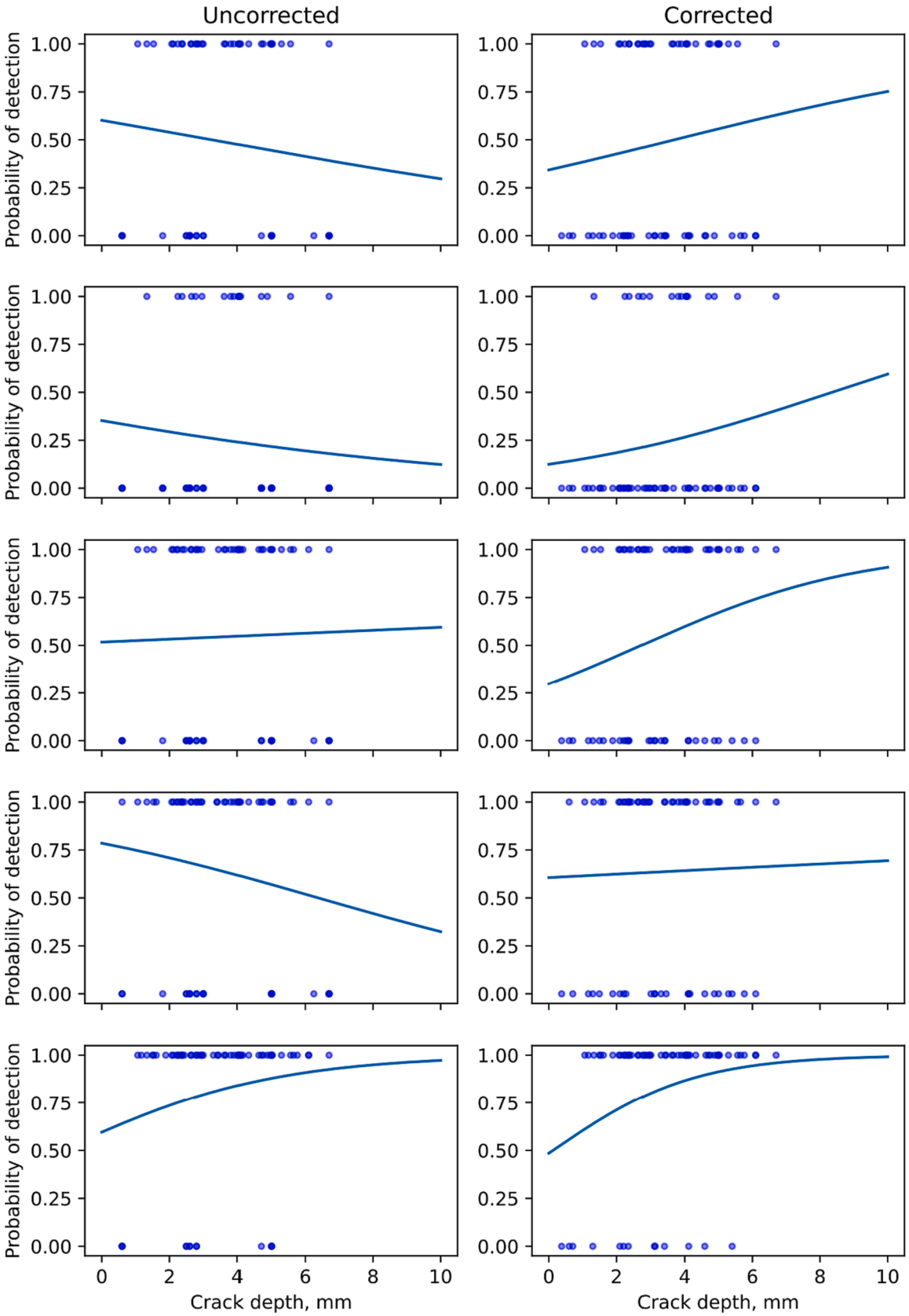

Fig. 10. Data set evaluation detection results as function of effective flaw size.

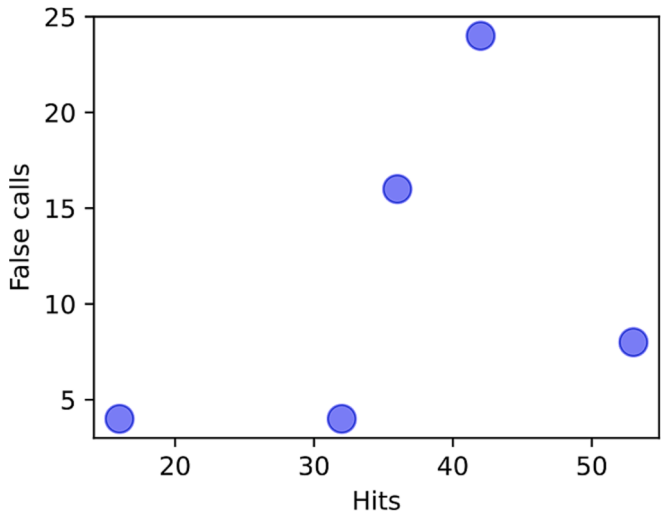

Fig. 11. False call count as a function of hit count. There's no apparent correlation between the detection performance and false call rate.

optimized for this procedure and thus should not be taken as an indication of ML performance, they do clearly indicate that automated data evaluation, when used to assist the human evaluator, can effectively address the challenges imposed by the need for rich data and allow the inspectors to focus on the most important data.

The sizing performance was low beyond what can be reasonably explained. Even accounting for the small number of result sets, the consistent noncorrelation of the estimates indicates highlights the difficulty of flaw sizing for such inspections. To some extent, this can be attributed to the microstructural effects described above and might perhaps be alleviated to some extent by richer data. However, this is clearly an important topic for further study.

5. Conclusions

The number of result sets in this virtual round robin was small: 5 sets for detection and of those 3 included sizing results. Despite the small number of results, the study revealed number of interesting features that were consistently exhibited by all the result sets.

The following conclusions can be drawn from the study:

- Increasing the data quality to modern multi-channel UT data is necessary to obtain realistic response, but also significantly increases the burden on data collection, virtual data preparation and data distribution.
- Increased flaw-free canvas data and latest generation of eFlaw process offer very high-quality data for this round robin.
- Consistent with previous studies, the detection performance displays significant variation between inspectors
- All the inspectors were liable to miss clear indications, when they were present in only few channels and, in particular, when the indications were in peripheral channels

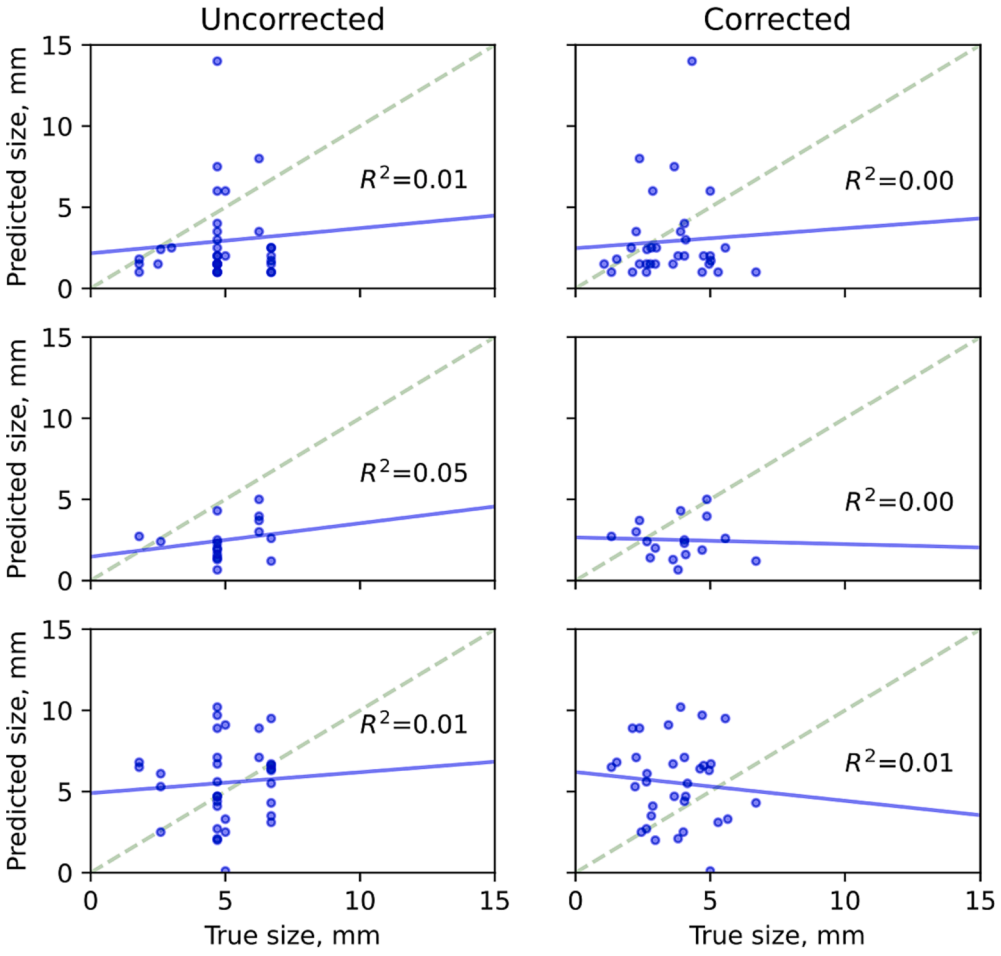

Fig. 12. Data set evaluation sizing results. Green dashed line shows the expected line, where predictions match true line. Blue line shows linear fit; in all cases the correlation is poor. The uncorrected plots (left) show crack sizes at their nominal size and the corrected plots (right) show sizes corrected with the applied amplitude factor.

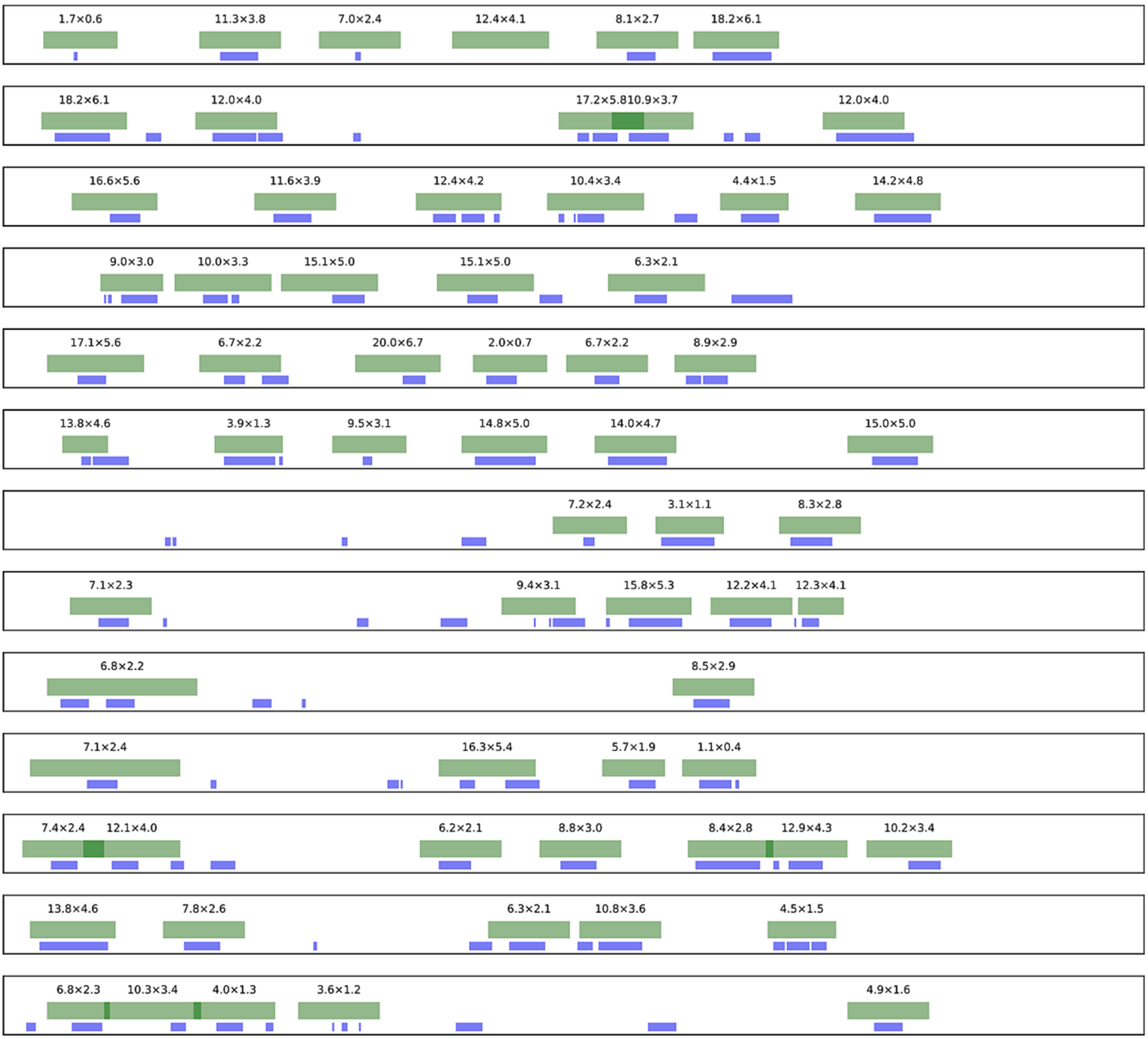

Fig. 13. Machine learning model detection results. The flaws in data files are shown with thick green bars and inspector indications with thinner blue bars below. The flaws are annotated with nominal size (size of the source flaw corrected with amplitude factor), length × depth, in mm.

- AI shows clear potential to alleviate these issues and allow more effective use of rich data sets
- Flaw sizing was unsuccessful in all the result sets provides and would merit further research

CRediT authorship contribution statement

Iikka Virkkunen: Conceptualization, Methodology, Validation, Formal analysis, Visualization, Writing – original draft. **Tuomas Koskinen:** Data curation, Investigation, Writing – review & editing. **Oskar Siljama:** Software, Investigation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The use of EPRI/Trueflaw DMW machine learning model for this study was supported by Thiago Seuaciuc-Osorio / EPRI. In addition, Trueflaw contributed the virtual flaw data generation and additional implanted flaws. Their support is gratefully acknowledged.

The research was funded by The Finnish Research Programme on Nuclear Power Plant Safety 2019 – 2022 (SAFIR-2022).

References

Becker FL, Latiolais, Carl, Bouck B. Dissimilar Metal Piping Weld Examination Guidance Volume 2. 2004;1009590.

Bevan, R.L.T., Croxford, A.J., 2022. Automated detection and characterisation of defects from multiview ultrasonic imaging. NDT and E Int. 128.
Cantero-Chinchilla, S., Wilcox, P.D., Croxford, A.J., 2022. A deep learning based methodology for artefact identification and suppression with application to ultrasonic images. NDT and E Int. 126.
Cantero-Chinchilla, S., Wilcox, P.D., Croxford, A.J., 2022. Deep learning in automated ultrasonic NDE – developments, axioms and opportunities. NDT and E Int. 131.
Crutzen, S., Lemaitre, P., Bièth, M., 1998. General lessons learnt from round robin tests and qualification tests relating to non-destructive examination. Int. J. Press. Vessel. Pip. 75, 417–427.
Doctor, S.R., Lemaitre, P., Crutzen, S., 1995. Austenitic steel piping testing exercises in PISC. Nucl. Eng. Des. 157 (1-2), 231–244.
Gantala, T., Balasubramaniam, K., 2021. Automated defect recognition for welds using simulation assisted TFM imaging with artificial intelligence. J. Nondestr. Eval. 40 (1).
Jacob RE, Moran TL, Holmes AE, Diaz AA, Prowant MS. Interim Analysis of the EPRI CASS Round Robin Study. 2018;PNNL-27712.
Koskinen, T., Virkkunen, I., Papula, S., Sarikka, T., Haapalainen, J., 2018. Producing a POD curve with emulated signal response data. Insight - Non-Destructive Testing and Condition Monitoring. 60 (1), 42–48.
Koskinen, T., Virkkunen, I., Siljama, O., Jessen-Juhler, O., 2021. The effect of different flaw data to machine learning powered ultrasonic inspection. J. Nondestr. Eval. 40 (1).
Kull, D., 2018. Nondestructive evaluation: cast austenitic stainless steel round-robin study. Tech. Rep. 3002010314.
Lemaitre, P., Koblé, T.D., Doctor, S.R., 1996. Summary of the PISC round robin results on wrought and cast austenitic steel weldments, part I: wrought-to-wrought capability study. Int. J. Press. Vessel. Pip. 69 (1), 5–19.
Lemaitre, P., Koblé, T.D., Doctor, S.R., 1996. Summary of the PISC round robin results on wrought and cast austenitic steel weldments part III: cast-to-cast capability study. Int. J. Press. Vessel. Pip. 69 (1), 33–44.
Lemaitre, P., Koblé, T.D., Doctor, S.R., 1996. Summary of the PISC round robin results on wrought and cast austenitic steel weldments, part II: wrought-to-cast capability study. Int. J. Press. Vessel. Pip. 69, 21–32.
Liu, T., Xia, S., Shoji, T., 2021. Intergranular stress corrosion cracking in simulated BWR water of 316L stainless steels manufactured with different procedures. Corros. Sci. 183.
López, H.F., Cisneros, M.M., Mancha, H., García, O., Pérez, M.J., 2006. Grain size effects on the SCC susceptibility of a nitrogen steel in hot NaCl solutions. Corros. Sci. 48 (4), 913–924.
Marois D. Procedure for Encoded, Phased Array Ultrasonic Examination of Dissimilar Metal Piping Welds. 2011;C3467_Zetec_OmniScanPA_03_revA.
Meyer R, Virkkunen I, Holmes A, Morales R, Seuaciuc-Osorio T, Lin B, Results of a Virtual Round Robin Study to Estimate Probability of Detection for Dissimilar Metal Welds 48th Annual Review of Progress in Quantitative Nondestructive Evaluation. 2021.
Posilović, L., Medak, D., Milković, F., Subašić, M., Budimir, M., Lončarić, S., 2022. Deep learning-based anomaly detection from ultrasonic images. Ultrasonics 124.
Reale, S., Tognarelli, L., 1995. Structural integrity significance of round robin testing trials, application to PISC III Action 3. Nucl. Eng. Des. 157 (1-2), 257–268.
Reale, S., Tognarelli, L., Crutzen, S., 1995. The use of fracture mechanics methodologies for NDT resuls evaluation and coparison. Nucl. Eng. Des. 158, 397–407.
Shipway, N.J., Huthwaite, P., Lowe, M.J.S., Barden, T.J., 2021. Using ResNets to perform automated defect detection for fluorescent penetrant inspection. NDT and E Int. 119.
Siljama, O., Koskinen, T., Jessen-Juhler, O., Virkkunen, I., 2021. Automated flaw detection in multi-channel phased array ultrasonic data using machine learning. J. Nondestr. Eval. 40 (3).
Svahn, P.-H., Virkkunen, M., Zettervall, T., Snögren, D., 2018. The use of virtual flaws to increase flexibility of qualification. 12th European Conference on Non-Destructive Testing.
Taheri, H., Gonzalez Bocanegra, M., Taheri, M., 2022. Artificial Intelligence, machine learning and smart technologies for nondestructive evaluation. Sensors (Basel) 22 (11), 4055.
Temple, J.G., Warm, J.S., Dember, W.N., Jones, K.S., LaGrange, C.M., Matthews, G., 2000. The effects of signal salience and caffeine on performance, workload, and stress in an abbreviated vigilance task. Hum. Factors 42 (2), 183–194.
Tyystjärvi, T., Virkkunen, I., Fridolf, P., Rosell, A., Barsoum, Z., 2022. Automated defect detection in digital radiography of aerospace welds using deep learning. Welding in the World. 66 (4), 643–671.
Virkkunen M, Rönneteg U, Grybäck T, Emilsson G, Miettinen K, Feasibility Study of Using eFlaws on Qualification of Nuclear Spent Fuel Disposal Canister Inspection. 12th International Conference on NDE in Relation to Structural Integrity For Nuclear and Pressurized Components, Dubrovnik, Kroatia. 2016.
Virkkunen, I., Miettinen, K., Packalén, T., 2014. Virtual flaws for NDE training and qualification. 11th European Conference on Non-Destructive Testing.
Virkkunen I. The “Small crack problem” in hit/miss Probability of Detection. NDT & E International. 2021;in review.
Virkkunen, I., Koskinen, T., Jessen-Juhler, O., Rinta-aho, J., 2021a. Augmented ultrasonic data for machine learning. J. Nondestr. Eval. 40 (1).
Virkkunen, I., Koskinen, T., Jessen-Juhler, O., 2021b. Virtual round robin–a new opportunity to study NDT reliability. Nucl. Eng. Des. 380.
Ye, J., Toyama, N., 2021. Benchmarking deep learning models for automatic ultrasonic imaging inspection. IEEE Access 9, 36986–36994.