Journal of Nondestructive Evaluation (2021) 40:67
https://doi.org/10.1007/s10921-021-00796-4

Automated Flaw Detection in Multi-channel Phased Array Ultrasonic Data Using Machine Learning

Oskar Siljama[1] · Tuomas Koskinen[2] · Oskari Jessen-Juhler[2] · Iikka Virkkunen[1]

Received: 18 August 2020 / Accepted: 29 June 2021 / Published online: 3 August 2021

Abstract
Modern ultrasonic inspections utilize ever-richer data-sets made possible by phased array equipment. A typical inspection may include tens of channels with different refraction angle, that are acquired at high speed. These rich data sets allow highly reliable and efficient inspection in complex cases, such as dissimilar metal or austenitic stainless steel welds. The rich data sets allow human inspectors to detect cracks with low signal-to-noise ratio from the wider signal patterns. There's a clear trend in the industry to even richer data sets with full matrix capture (FMC) and related techniques. Convolutional neural networks have recently shown capability to detect flaws with human level accuracy in ultrasonic signals at the B-scan level. To enable automated flaw detection at human-level accuracy for critical applications, these neural networks need be developed to take advantage of today's rich phased array data-sets. In the present paper, we extend previous work and develop convolutional neural networks that perform highly reliable flaw detection on typical multi-channel phased array data on austenitic welds. The results show, that the modern neural networks can accommodate the rich ultrasonic data and display high flaw detection performance.

Keywords Machine learning · NDT · Phased array · Image classification

1 Introduction

Conventional ultrasonic weld inspection requires multiple physical probes with different angles to achieve satisfying results. Phased array systems can be used to reduce the amount of required probes. Phased array probes consist of a transducer system of multiple elements that can be controlled, pulsed and received, separately. By controlling the transducer elements through focal laws one probe can be used to produce different beam angles, beam steering and focus depths. For weld inspections, mechanised scanning allows for the inspection data to be recorded consistently and more importantly allow more thorough data analysis possibilities afterwards. Although this is also possible for conventional probes, it requires multiple individual scans and probe angles making it more time consuming. A system of multiple conventional probes could be used to limit the amount of required scans, but such systems are too large in size to be utilized in inspections. Phased array systems limit the amount of required scans with transducers small enough for inspection. While phased array system have been around for medical systems for decades, they have become more common in NDT in the late 00's and their use has been steadily increasing [5,10,12,30].

Developments in the phased array systems allow increasing amounts of data to be captured in the inspections with similar time frame. This means systems like full matrix capture (FMC) or total focusing method (TFM), which record every transmit-receive combination possible by the transducer producing a fully focused and comprehensible image, have become available. Modern phased array system allow data interpretation in a way that is more generally understandable [17,18,30]. Using the same data format for the inspectors and machine learning (ML) models enables direct comparison between the ML system and an inspector.

In nuclear power plants (NPPs) NDT methods are used to verify the structural integrity of critical components. Inspections are carried out during maintenance outages making the access to the inspection targets limited. Mechanical scan-

✉ Tuomas Koskinen
tuomas.koskinen@vtt.fi

Oskar Siljama
oskar.siljama@aalto.fi

[1] Aalto University, PL 14200, 00076 Aalto, Finland

[2] VTT Research Centre of Finland, Espoo, Finland

ning systems are used to gather ultrasonic inspection data from primary circuit components reliably and consistently. After monitoring the data collection the inspector will go through the ultrasonic data and look for flaw indications. This data analysis is a tedious and time consuming task with a possibility to human errors. According to Ali et al. [3] most of the human errors are caused by poor instructions and fatigue. Bato et al. [4] studied the environmental factors on human inspectors and humans performed worse in the field than in the laboratory. Moreover, there are differences on the performance of different inspectors, conducting the same inspection [33,39]. Automated systems are not susceptible to such errors. Thus, automation has been used for ultrasonic testing for decades in areas where data interpretation is clear and the amount has been high [16]. The major drawback of traditional automated systems has been the inability to handle noisy and complex ultrasonic data.

ML models have proven their effectiveness in various image recognition tasks Aggarwal [1], Chowdhury et al. [9], Munir et al. [25,26], Virkkunen et al. [40] thus ML models could be employed to remove the bulk of the repetitiveness of NDT data analysis, even in the noisy and complex cases. Since the majority of the inspection data is usually without flaws, the ML model could be used to look for flawed areas. After identifying the locations of the possible flaw indications via a ML system, the inspector could verify the results and apply expert judgement in flaw evaluation. In high reliability industries like the nuclear industry the use of best practices is mandated. The ability to utilize increasing amount of inspection data allows for flaw detection at an earlier stage as well as more effective monitoring of the system and the flaws.

The ML approaches for image classification have been also developed in increasing speed. State of the art DCNN for image classification tasks can be considered the YOLO networks. YOLO networks. [32] introduced a very deep convolutional network VGG, comprising of 16 to 19 weight layers with small convolutional filters suited for large-scale image recognition tasks with vectorized output for classification. The VGG network performed outstandingly well as it achieved first and second place in ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC-2014) image recognition and localization tasks, respectively. Redmon and Farhadi [27,28] introduced the YOLO deep neural network for object detection tasks, with very high speed real-time image processing that is continuously improving. Compared to VGG, YOLO outputs a tensor representing a grid of the input image, containing both classification labels and coordinates for the bounding boxes for each grid.

The recent developments in machine learning have also found their way into non-destructive testing (NDT). The areas where the NDT data is natively image-like can often take almost direct advantage of the recent advances in ML for image detection. These areas include, visual testing Chen and Jahanshahi [6,7], Tang and Chen [35], radiographic Du et al. [14] testing and eddy current testing Zhu et al. [42].

ML powered ultrasonic inspection operates under the same principles and constraints as any other form of machine vision and image recognition. However, modern ultrasonic inspection differs slightly from the traditional image recognition case as data gathering enables large variety of different options in terms of angles, focus and scan plans. Hence, the same scan location produces multiple images and depending on the index location, some parts of the scanned information are essential for flaw detection while other parts are completely irrelevant. This increased amount of images might needlessy consume computing resources or lead to missing of the flaws if scanned data would need to be constrained. In this paper, we study how a machine learning model can handle ultrasonic phased array data and how this type of data should be handled most efficiently and fully utilizing the gathered scan data. In addition, the model is trained with flaws that are scanned in base material and virtually implanted to the weld scan data. The model performance is tested with real thermal fatigue cracks in similar weld geometries as used in training and compared to the results of a human inspector.

1.1 Machine Learning for UT Data

Ultrasonic data has been interpreted before with simple and shallow neural networks, usually for A-scan classification. [8,22,41] Support-vector machines (SVMs) have been an effective way to classify ultrasonic signals as demonstrated by Fei et al. [15], Matz et al. [23]. For both the shallow neural networks and SVMs have reported high classification accuracy for A-scan, these methods have been unfeasible for wider use due to feature engineering. In feature engineering the models require feature extraction by hand, which deteriorates the scalability. While the number of extracted features were only 12 for Sambath et al. [29] and 5 for Cruz et al. [11] using SVMs, the deep neural networks can be trained with much higher efficiency as the feature engineering is left for the model.

More recent approaches have been used with the PAUT data. Luo et al. [20] were able to construct an algorithm which utilized spatial clustering and segmentation to detect flaws from the S-scan data in TKY welded joints. While the algorithm used multiple angles, the algorithm relied only to the 2D data obatained from the S-scan. Shukla et al. [31] utilized a physics-informed neural network to detect surface breaking cracks in ultrasonic data. The model based the detection on the decrease in the speed of sound recorded from the signal.

The DCNN for NDT purposes have been shallow when comparing to the modern deep networks. The DCNN Virkkunen et al. [40] utilized had less than 100,000 parameters. The ultrasonic data can be considered simple as there are a lot of similarities within the data, which enable the use of simpler mathematical approaches as for Luo et al. [20] to a certain point. Moreover, the number of required extracted features for high classification accuracy has been low for shallower networks and SVMs. Thus, the shallower DCNN and simpler approaches have been successful in classification tasks.

The recent advances in convolutional networks have enabled the use of machine learning also for automated flaw detection in complex UT signals—a task that was considered infeasible for a long time. Munir et al. [25,26] used convolutional neural networks to analyze single A-scans and reported impressive results for flaw detection in weldments. However, the information contained within a single A-scan is limited and also human inspectors typically move the probe interactively during inspection to sample multiple A-scans and to get information from the echo dynamics of the reflector. For mechanized inspection, this probe movement is done by mechanical or electronic (phased array) scanning, and the information can be obtained by analyzing a set of related (adjacent) A-scans, i.e. B-scans. Virkkunen et al. [40] used convolutional networks to analyze B-scan level data and obtained human-level performance in flaw detection. For many NDT systems, there is a requirement to use best-available techniques and so obtaining human-level performance is a significant milestone that enables wider application of ML systems in practice.

Despite these advances, the ever richer data sets acquired in modern phased array UT inspections provide a significant challenge for the ML networks. As the data size increases, the computational burden of training also rapidly increases. Thus, the key challenge remains to adapt the rich UT signals so, that the computational burden remains feasible without significant loss of signal data.

1.2 Training Data for NDT

The use of richer ML networks for flaw detection and also the widening of the field-of-view from separate A-scans to B-scans and beyond is necessary to obtain human-level performance. However, it also significantly increases the amount of data needed for training. In NDT, we typically have copious data for un-flawed inspection but flawed data sets are scarce. The lack of representative flawed data sets is a common problem even for training humans, which require substantially less data than ML systems.

A common technique in ML to work with limited training data is to use data augmentation. By applying different transformations to the training data, variation that should be inconsequential to the classification but is not well represented in the training data can be artificially introduced and the models can be made to generalize much better. Traditionally, these augmentations include image manipulations like shear, rotation and scaling. For NDT data, the signals are simple, in comparison to typical image detection tasks. In contrast, the issue is typically to find signals with very low signal-to-noise ratio. Thus, the extent that such traditional data augmentation can provide improvement is limited [14].

A more sophisticated augmentation scheme can be obtained using the so called virtual flaws [34,36–38]. The central idea with the virtual flaws is, that the flaw signal can be extracted from the background and then re-introduced to different backgrounds and separately augmented to provide additional variation. This approach allows significantly richer augmented data. It has been successfully used in training and qualification of human inspectors and even in POD evaluation of humans in [39]. It's also been used for training machine learning networks for various techniques [40].

In addition to data augmentation, generating the data through simulation has been used for eddy current inspections. Miorelli et al. [24] used Output Space Filling (OSF) to generate teaching data and Ahmed et al. [2] used similar approach by adding Partial Leas Squares (PLS) feature extraction to OSF. These approaches were able to train machine learning models within efficient computational time. Ahmed et al. [2] speculated that this approach could be plausible for ultrasonic and thermographic methods as well.

2 Materials and Methods

2.1 Ultrasonic Set-Up

Ultrasonic data was acquired with the same procedure as used for normal pre-service ultrasonic inspection in NPPs and what is normally recommended for austenitic welds [13]. The original data was scanned using a dual matrix phased array probe with 2.25 MHz frequency in transmit-receive longitudinal (TRL) setup. There were total of 28 elements per probe with the arrangement of 7 × 4 elements and active aperture of 19 × 12 mm. The probes were placed in a rexolite wedge with an angle of 18.9° and 0° roof angle with 6 mm separation between the probes. Focal laws were set up for 40° to 70° angles with 1° step. Scan resolution was 1 mm and the sound path was set from 3.46 to 27.75 μs with sound path resolution of 0.01 μs, thus total data size for single scan step was 2429 × 31 samples. Focus point was set to the bottom of the weld, middle of the probe with no skew angle. The probe was positioned such that 55° angle would be centered at the weld root as only one scan line was recorded. The schematic for the probe and wedge setup can be seen in Fig. 1.

Probe movement was recorded by a single encoder while a single scan line was manually scanned along the weld. Probe deviation from the said line was prevented by a stationary rail in front of the scan line. Sensitivity calibration was done to a standard austenitic stainless steel block with 1 mm machined

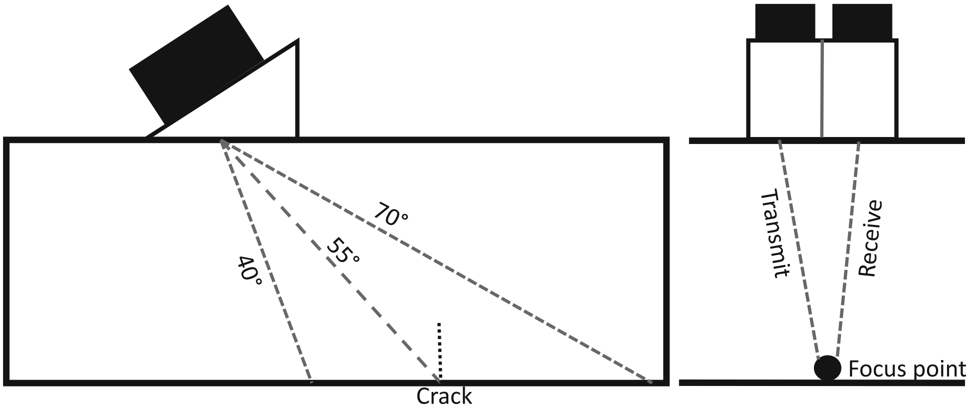


Fig. 1 Schematic of the probe and wedge setup. The phased array focal laws were focused to the bottom of the plate with azimuthal scan from 40° to 70°

notch. Gain was adjusted so the 55° angle would reach 80% of the maximum amplitude on the notch. The scan was made by a certified level 3 inspector with Zetec TOPAZ64 phased array equipment. Contact medium used was water applied by hand from a spray canister.

2.2 Physical Samples for ML Training

For the raw UT data, two set of simplified plate samples were used. Three plate samples with weld in the middle was scanned to provide flaw-free UT data. The samples differed in thickness and roughly represented the range of thicknesses expected in the final application. All the samples are shown in Table 1. The material for all the samples was AISI 316L stainless steel and welded with Gas tungsten arc welding (GTAW) or Shielded metal arc welding (SMAW) . The weld samples (1–3) were scanned from both sizes. In addition, separate flawed samples were scanned to acquire flawed data. The flawed samples did not contain any welds and thus provided noise free flaw signals. The flaw sizes are shown in Table 2. The flaw samples were scanned from both sides and with three scan distances each side (focused at root, +5 mm and +10 mm). As is characteristic of all natural flaws, the cracks exhibit variation as they find their way through the local microstructure. Thus, the flaws are expected to give somewhat different UT signal when scanned from different sides (and with different distance and angle).

This scan set-up maximized the amount information obtained from each sample and yielded 6 distinct flawless weld backgrounds and 80 distinct flaw signals from 16 physical flaws.

The present set-up where un-flawed weld-data and scans from flaws introduced in base material provides a flexible arrangement and allows insertion of flaw signals to various locations in the weld. Some of the flaws were so small as to be at the limit of detection, if manufactured directly in the noisy weld. The present set-up allows extraction of clean flaw signal from relatively low-noise samples and embedding it into the noisy environment where detection is uncertain or even impossible.

Table 1 The scanned samples included flaw-free weld samples and flawed base material samples detailed here

#	Type	Thickness	Notes
1	Weld	30 mm	Mechanized SMAW
2	Weld	30 mm	Narrow-gap GTAW
3	Weld	20 mm	Narrow-gap GTAW
4	Base material	20 mm	Flaw 1
5	Base material	20 mm	Flaw 2
6	Base material	20 mm	Flaw 3
7	Base material	20 mm	Flaw 4
8	Base material	20 mm	Flaw 5
9	Base material	20 mm	Flaw 6
10	Base material	20 mm	Flaw 7
11	Base material	20 mm	Flaw 8
12	Base material	20 mm	Flaw 9
13	Base material	20 mm	Flaws 10–12
14	Base material	20 mm	Flaws 13–16

While the set up has significant advantages, it also misses the possible interaction with weld microstructure and flaw growth. It is conceivable, that flaws grown directly to, e.g., weld fusion line, would be affected by the local microstructure and display different characteristics than flaws grown in base materials. Previous studies by Svahn et al. [34] indicate, that such variation is negligible for the present case of mechanized butt-welds in austenitic stainless steel. Thus, separate final validation flaws were manufactured with flaws produced directly into the weld fusion line.

2.3 Physical Samples for Validation

A completely separate sample set was created for the final validation of the trained machine learning network and to evaluate its performance against human inspectors. In con-

Table 2 The flawed samples included altogether 16 defects detailed in the following table

#	Flaw id	Length (mm)	Depth (mm)
1	058CAB6775	1.6	0.6
2	060CAB6780	2.3	0.9
3	061CAB6781	3.5	1.3
4	045CAB6750	3.9	1.4
5	057CAB6774	6.0	1.4
6	160BCB1672	3.9	1.4
7	185BCB1709	3.5	1.8
8	220BAB1152	3.7	1.8
9	049CAB6759	5.4	2.0
10	045CAB6749	8.8	3.3
11	230BCB1740	10.3	4.9
12	013BFB2687	13.8	5.9
13	222BAB1153	24.2	6.8
14	251BBB1399	23.9	7.1
15	188BCB1715	28.4	7.6
16	173BCB1697	18.9	9.0

All defects were true thermal fatigue cracks, grown in-situ to the base material samples (i.e. not weld implanted) by Trueflaw Ltd. The manufacturer's flaw id is provided for traceability

Table 3 The scanned validation samples included flawed weld samples

#	Type	Thickness	Notes
1	W2685	20 mm	Narrow-gap GTAW
2	W2686	20 mm	Narrow-gap GTAW
3	W2690	30 mm	Mechanized SMAW

trast to the samples used in training the network, in this validation set real flaws were manufactured directly to the interesting weld fusion line locations. The sample geometries similar to those described in Sect. 2.2 and are listed in Table 3. All defects were in-situ produced thermal fatigue flaws manufactured by Trueflaw Ltd. The validation flaw sizes are shown in Table 4. The true depth of these flaws is not validated, and can be roughly estimated from the surface length and the expected aspect ratio. The validation samples were scanned from both sides yielding 6 data files, with 16 theoretical flaw indications with 8 on the near-side (easier) and 8 in the far side (more difficult). The data for the separate validation samples was gathered on a separate occasion, on a separate (but compatible) machine and by a different inspector, following the same procedure as the training data. This disconnection between the data gathering is expected to minimize any possibility for overly-homogeneous training/validation data that would limit generalisation of the model on later acquired data.

Table 4 The flaws in the validation set included altogether 8 flaws

#	Flaw id	Length (mm)	Depth (mm)
1	346CAB7300	7.5	(2.0)
2	346CAB7301	11.5	(3.5)
3	349CAB7306	5.0	(1.5)
4	346CAB7302	9.0	(2.5)
5	346CAB7303	9.0	(2.5)
6	349CAB7307	5.0	(1.5)
7	349CAB7308	5.0	(1.5)
8	350CAB7315	3.5	(1.0)

All defects were true thermal fatigue cracks, grown in-situ to the weld fusion line (i.e. not weld implanted) by Trueflaw Ltd. The manufacturer's flaw id is provided for traceability. The depth values are not validated, and should be considered rough estimates

The criterion for ML model was to find flaws as small as possible, while avoiding false calls in the process. For human inspector the instructions were similar and to follow traditional inspection protocol for austenitic stainless steel welds. Thus, no hard amplitude limit was set for detection, but the inspector looked for deviations in the geometrical indications and to detect unique signal inconsistencies from the data. The inspector was highly familiar with austenitic weld inspection.

2.4 Data Development and Augmentation

The flaw signals were extracted from the acquired data files to facilitate eFlaw data augmentation With 16 cracks and 5 scans for each crack, the raw data contained altogether 80 extracted flaw signals. Similarly, the multiple scans of welds resulted in altogether 6 flawless canvases for data augmentation, one of which was designated unusable due to differing acquisition set-up. Since the flawed samples did not contain the characteristic weld noise, they could not be used as augmentation canvases.

Moreover, the signal was gated for the longitudinal wave and the shear wave component was left out from the signal shown to the model.

The designated unit of analysis was a partial scan with 48 A-scans clipped to 1020 samples from the interesting weld region for each of the 31 channels. The clipping to 1020 samples left only the longitudinal component in the data. While the shear wave component can be used for flaw detection in some cases, for these austenitic welds the noise was so high the focus on just longitudinal component was justified. Thus, a single data sample was $48 \times 1020 \times 31 \approx 1.5M$ samples in size.

A balanced, augmented data set was generated as follows:

1. Each weld file was used as a flawless canvas. The full acquired A-scans were clipped to the interesting region around the weld to minimize excessive data.
2. For each canvas, a set of 500,000 samples were generated (divided in 50 batches of 10,000 samples each):
   (a) A random number was picked to select flawless or flawed sample.
   (b) If the sample was designated as non-flawed:
      i. random window of 48 A-scans was selected and added to the result data.
      ii. The data was further augmented by shifting each A-scan by a random-walk offset that mimics possible probe jitter during scanning.
   (c) If the sample was designated as flawed:
      i. A random flaw was picked from population and embedded to random location in the file.
      ii. The flaw amplitude was decreased by random factor in the range 0.5…1.0.
      iii. The flaw was augmented by shifting each A-scan by a random-walk offset that mimics possible probe jitter during scanning.
      iv. After embedding, a random window of 48 scans were selected such, that the flaw was wholly contained within the window.
      v. The data was further augmented by shifting each A-scan by a random-walk offset that mimics possible probe jitter during scanning.

In addition to the raw data, the corresponding labeling data was recorded as a text file, that included the flaw location rectangle in the sample, flaw original size, computed equivalent size (based on the amplitude fraction) and location in the original canvas.

Data augmentation through virtual flaws resulted in altogether of 500,000 samples with approximately 50% flawed. The data was stored in compressed binary form and took roughly 500 GB of storage. For efficiency, the data preprocessing described in Sect. 2.5 was integrated in the data augmentation. Thus in addition to the full raw embedded data, a reduced preprocessed data was generated. The preprocessed data was used for the actual training, while the full data was used to confirm data quality and preprocessing efficiency.

The available flaws and canvases were divided between and training, validation and test set (excluding the separate samples for final validation). A separate validation set was also used during development, but due to limited number of flaw-free canvases, the separate validation set was reduced the training set excessively. Thus the final models were trained with all the available non-test flaws and backgrounds. Due to these reasons, validation set containing 10,000 samples was separated from the training data and the validation set contained the same physical flaws and backgrounds that were used in training. The physical flaws or canvases selected for the test set were excluded from the training/validation set and the reported performance is achieved from the test set, which contained a total of 1000 samples. All the data sets contained 50% flawed images and 50% un-flawed images.

In addition, the smaller flaws (1–10) proved impossible to discern when embedded in the weld noise. Thus, the training data set was also divided to "big" and "all" flaws and training was tried with both. The test set included the designated small flaws in both cases.

2.5 Machine Learning Model

Before the data was submitted to the model, it was preprocessed in a following way. The multiple angle channels contain a significant amount of redundant data. Furthermore, significant portion of the A-scan samples are needed to acquire the waveform and to avoid anti-aliasing effects with the ultrasonic waveform. However, the signal is already significantly frequency filtered due to resonance of the used probes (often accompanied by further electronic filtering to reduce noise). Thus, the frequency data contains little significant information for flaw detection for this case. Human inspectors also do not usually make use of the phase or frequency information: they typically look at the data rectified and merged, which covers any phase and frequency related information in the raw data. Thus, the ML process can be made more efficient by preprocessing the data in a way that reduces sampling while still retaining the significant signal information. In principle, the pre-processing could have been implemented as part of the neural network as a trainable layer, however separating the simple pre-processing step improved the data augmentation efficiency significantly. Various preprocessing methods were tried and the following preprocessing pipeline was chosen:

1. Each of the 31 full waveform channels are considered separately
2. Eeach frame is rectified (i.e. absolute value of the signal taken).
3. The frame is max-pooled with window that matches the $\frac{1}{2}\lambda$. This has the effect of taking a computationally efficient envelope of the data. The data size is reduced from 48×1020 to 48×34 (=1632) samples.

This preprocessing enhanced the data efficiency of the process significantly and also facilitated generalization of the model, since small differences in scan parameters (number of channels, channel angles, A-scan lenght, Scan length) can be accommodated in the preprocessing stage. The data was then stored to compressed binary files to facilitate file transfer and accelerate learning. For training, the data was decompressed,

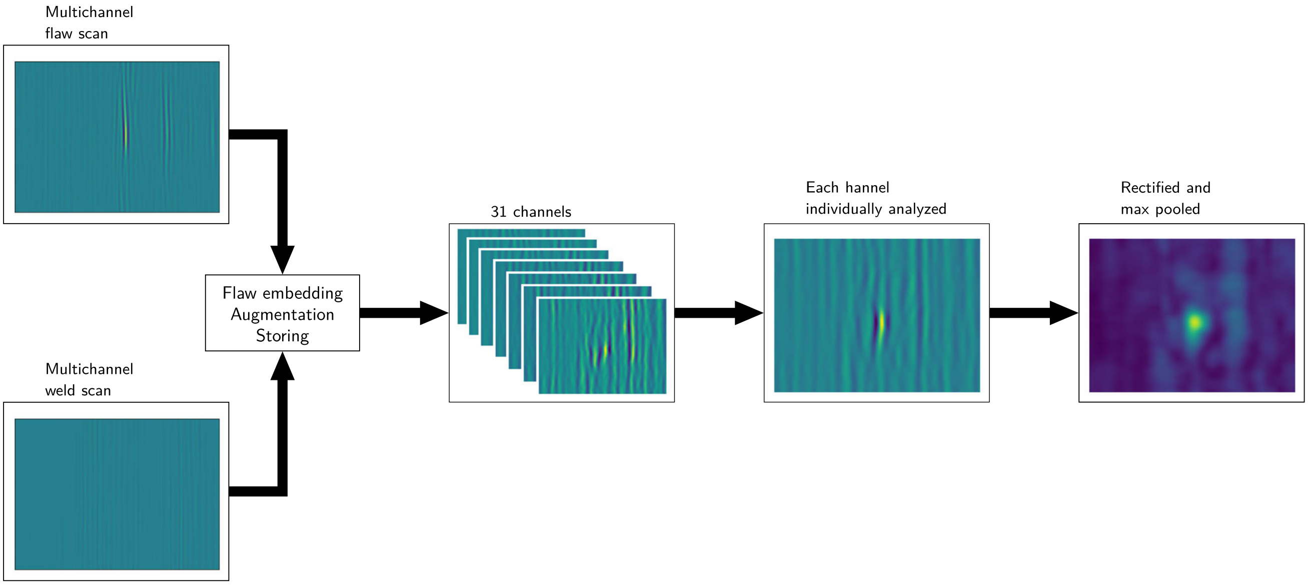


Fig. 2 Preprocessing pipeline of the ultrasonic data. First the data from homogenous material, containing the flaw data, and the weld is obtained. The flaw was then embedded to the weld and necessary augmentation was conducted to enrichen the data set. Each channel of the sample was then clipped to the area of interest. The single channels were then rectified and max pooled according to the $\frac{1}{2}\lambda$ and evaluated individually

and converted from the original 16 bit integers to 32 bit floating point numbers and scaled to 0…2.0 (with most data in range 0…1.0).

In addition to the selected model, the following alternate schemes were tried:

1. Full 31 channels used as input to the network (rectified and max-pooled)
2. 31 channels summed and used as a single channel, then (rectified and max-pooled)

Network

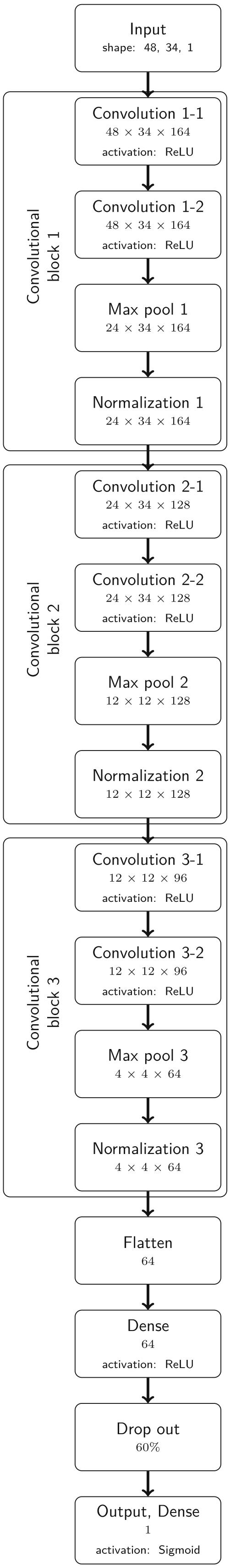
The used DCNN architecture resembles VGG16 [32] network with 3 convolutional blocks. Each block contained two consecutive convolution layers with rectified linear unit (ReLU) activations. This was followed by a batch normalization (BN) layer to normalize the input distribution to the following block to increase robustness of the network by reducing the internal covariate shift [19]. The convolutional blocks were followed by vectorization and a densely connected layer with ReLU activation and units corresponding to the number filters of the last convolution. Finally, the dense connected layers weights converged to a single classification unit with sigmoid activation, indicating if a crack is present. The loss function applied was binary cross-entropy. For computing the new weights during backpropagation, adaptive moment estimation (ADAM) was used. The used network is visualized in Fig. 3. Computations were conducted utilizing TensorFlow library for data flow in preprocessing and filtering, and Keras high-level API for constructing the DCNN.

2.6 Validation Data Evaluation

The data from the separate flawed validation samples was run through the trained machine learning model as follows: each file was split to a set of individually evaluated data-frames corresponding to the chosen trained model input data size (96 rows). The frames were construed by moving a window of the said size with 50% overlap throughout the data, i.e. the first frame contained lines 1–96, the second lines 49–144 and so forth. For each frame, all the 31 channels were separately evaluated. If any (even one) of these frames were designated as flawed, the frame location was considered as flawed. If a frame containing a flaw was identified as flawed, this was considered a true hit and if identified as un-flawed it was considered a miss. If a frame was indicated as flawed but did not contain a flaw, it was considered a false call. The data did not contain any cases, where a flaw would partially fall on a frame. Altogether, the data so divided contained 32 separate data frames with 11 opportunities for hit/miss on the near side, 11 opportunities for hit/miss on the far side and 10 opportunities for false calls.

The data files were also given to a human inspector for a blind evaluation. The inspector remained oblivious to the flaw location, but did have information regarding the pairiwise arrangement of the data files (i.e. new which files were acquired from different sides of the same weld and thus con-

Fig. 3 Used deep convolutional neural network for estimating flaws in ultrasonic scans



tained the same flaws). This may have helped the inspector to identify unclear flaws from the far side.

3 Results

3.1 Ultrasonic Results

The austenitic stainless steel blocks scanned with flaws produced a clear ultrasonic image. Even the smaller flaws were easily detectable as the material is low noise and homogeneous. The empty weld canvas represents a typical austenitic stainless steel weld. Since the base material is homogeneous the section is easily interpreted. However, the anisotropic weld produces a lot of noise and attenuates the sound considerably. When the flaws were implanted to the austenitic weld through the eFlaw process it rendered the smallest flaws virtually undetectable. Even some of the flaws in the medium size range were difficult to detect if they vere implanted on an especially noisy location. In general this shows that the welds behave as expected and the eFlaw flaw implantation from a surrogate flaw sample has worked as expected.

3.2 ML Performance

Initial ML model performance was measured by estimating previously unseen testing data set, extracted from the data set containing all the available flaw sizes, with roughly 50 % scans with cracks and 50 % without to measure the true performance of the model and observe possible overfitting. The results were evaluated based on false call rate and probability of detection (POD) metrics, which are also used to evaluate human inspectors. With data augmentation, the number of data points in the POD computation is very high (500) compared to a traditional POD exercise (typically around 60). However, the augmented flawed images are, of course, much less independent than in a traditional POD exercise and display more limited variability than actual inspection data would.

From the three channel combination modes, shown in Sect. 2.5, the other modes tested exhibited some adverse behaviours described in the following. Using the full 31 channels was expected perform well, but it had severe overfitting issues that proved unresolvable. These may be caused by the limited amount of different flaw-free canvases, that allowed the network to preferentially memorize the flaw-free canvases. Thus, with additional variation in the free canvas data, the full 31 channel model may prove superior. When all 31 channels were pre-combined before ML-evaluation, the overall results were good, but showed persistent misses of unacceptably big cracks. It appears that for some small percentage of the cracks, the combination tended to conceal the crack signals.

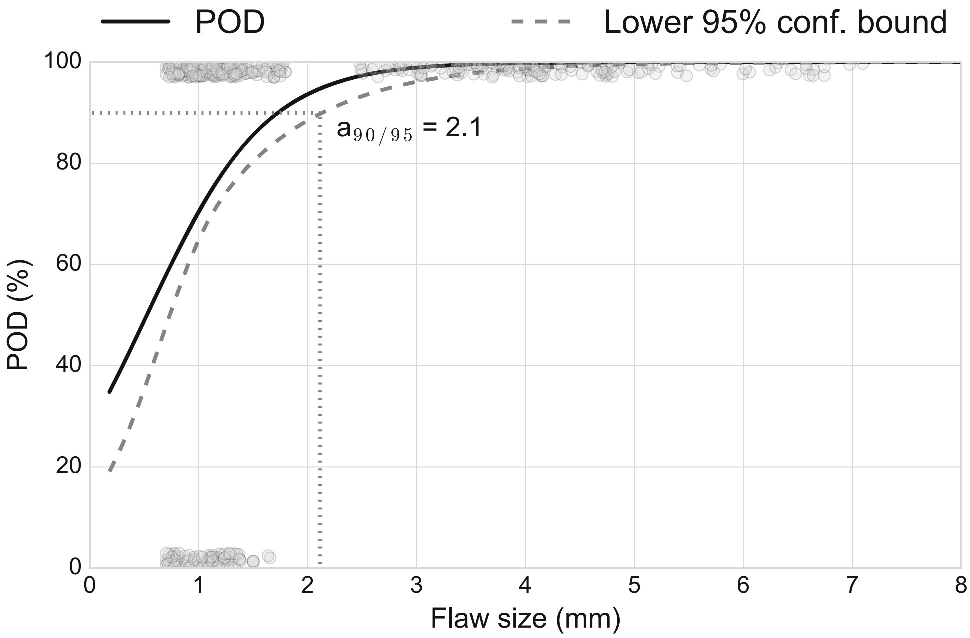


Fig. 4 POD curve for the model trained only with the bigger flaws. The small flaws in the POD estimation were completely new for the model. Flaw size represents the flaw depth in mm. Solid black line represents the POD value and the dashed grey line the lower 95% confidence bound, $a_{90/95}$ = 2.1 mm with false call rate of 2.3%

When the model was trained with all the cracks, including the small indication that are expected to be impossible to find, the model showed excessive false call rate of 14%. When the model was trained with bigger cracks only, the model exhibited a better false call rate of 2.3%. When the smallest cracks were removed, the model found all the big cracks and also indicated some small cracks to offer quite good POD values ($a_{90/95}$=2.1 mm). Thus, the models showed good generalization, as the small cracks were completely new flaw type for the model. The achieved POD curve can be seen in Fig. 4.

The final evaluation was completed with the separate final validation data described in Sect. 2.3. The model performed consistently with the training data: all $> 5mm$ surface length flaws were found from the near side, while all $< 5mm$ flaws were missed. This is consistent with the selected training flaws. From the far side, the model found 3 out of the 4 $> 5mm$ flaws, but missed one 9 mm flaw (flaw 5 in Table 4, designation 346CAB7303). The far side scan image showing the missed flaw 346CAB7303 and detected flaw 346CAB7302 can be seen in Fig. 5. Both of the flaw indications are difficult to distinguish from the noise. The flaws from the far side were detected with different channels (larger angles) than the near side, and showed markedly less salient indications. Also, the far side indications exhibited larger variability due to the sound path passing through the inhomogeneous weld, which explains why one of the cracks were missed as the flaw was situated in the middle of the flaw in the training data and not on the far side. While the missed crack was not the smallest, it was the one of the smallest indications. The model made two false calls, both corresponding to somewhat crack-like signal pattern associated with scanning extending over the sample end; a condition which was not included in the original training data due to different sample arrangement. The human inspector showed performance similar to the machine learning network: The human inspector found all the flaws the ML found, and in addition found one flaw from both sides whereas the ML only found it from the other side. In addition, the human found one very small crack, which can also be a lucky false call. The human inspector made four false calls, in total. The full performance comparison is shown in Table 5.

4 Discussion

As previously stated, the current ML models are rich enough to handle complex ultrasonic data and to reach very high detection capability on noisy data. The present results indicate, that the models also extend well to the rich multichannel data provided by the modern phased array ultrasonic equipment. The scan data was considerably noisy and it is clear no static amplitude threshould could have been used. Figure 6 highlights the areas which exceed noise threshold of 6 and 12 dB. A mean amplitude of the whole image was calculated and set as the noise level. Only the B7300 and B7301 could be detect with these thresholds, while making considerable amount of false calls in the process.

For the present case, the channels were considered individually and flaw indicated if any of the channels indicated a flaw. A priori, it was expected that training on the full data with all the 31 channels concurrently would provide better results, since it could learn to combine the information in the various channels. For the present data, this potential opportu-

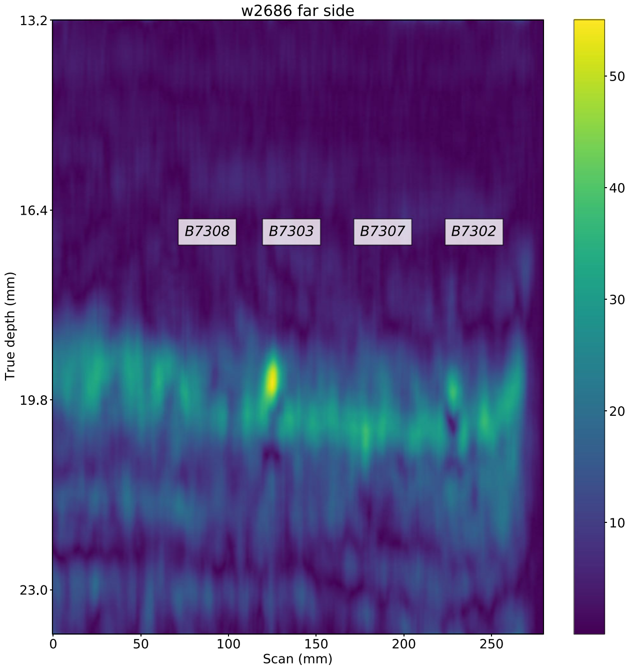


Fig. 5 Weld W2686 B-scan image at 59° from the far side and the flaw B7303 found by the human inspector but missed by the ML model

nity was shadowed by the tendency to overfit on the limited background data. In the future, the overfitting issue could be alleviated by increasing the amount of flawless data in the data set. Flawless weld data is normally available and so acquiring additional flawless data is not expected to be problematic. Similarly, although the achieved false call rate (2.3%) can be considered acceptable, as for human inspectors this can be around 1 - 9 % in noisy inspection cases Maier et al. [21], Virkkunen et al. [39] . However, this false call rate could be easily improved by increasing the amount of flawless data and variation within the flawless welds.

The present study indicated, that the selection of the training flaws has significant impact to the ML performance. In particular, including very difficult or impossible to find cracks can easily result in excessive false call rate, which is not unexpected. For this study, the small flaws were removed from the training set to avoid that issue, but were still detected to certain extent by the trained models. In many applications, the inspector is expected to follow a clear detection limit and to only report flaws above a certain size. From this perspective, the indicated small flaws could be seen as false calls (even if they are indications of a real flaw). In this case, the small flaws could be included in the training set as non-flaws to omit these indications. However, this was not done in the present study, since the primary interest was to find true limits of detection.

The present network may seem excessive for the fairly simple detection of UT indications in the pre-processed data. In fact, we tested with significantly lighter networks (as low as ≈ 60,000 parameters) and they exhibited good flaw detection accuracy. However, to reach stable and low false call rate, it was necessary to utilize a larger network. This effect may be related to the limited set of un-flawed canvases.

While the validation set with the human inspector was lacking the flaw amount for proper statistical evaluation, the result is highly promising. The model almost managed to match the inspecor’s performance, missing two cracks of which one might have been found due to extraordinary circumstances and the other due to lacking far side training data. Main reason for the ML model to miss the one larger far side

Table 5 Final validation blind result comparison from the machine learning network and from a human inspector

Sample	Flaw	Side	ML	Human	Notes
W2686	346CAB7302	Far side	Hit	Hit	
W2686	346CAB7303	Far side		Hit	
W2686	349CAB7307	Far side			
W2686	349CAB7308	Far side			
W2686	346CAB7302	Near side	Hit	Hit	
W2686	346CAB7303	Near side	Hit	Hit	
W2686	349CAB7307	Near side			
W2686	349CAB7308	Near side			
W2685	346CAB7300	Far side	Hit	Hit	
W2685	346CAB7301	Far side	Hit	Hit	
W2685	349CAB7306	Far side			
W2685	346CAB7300	Near side	Hit	Hit	
W2685	346CAB7301	Near side	Hit	Hit	
W2685	349CAB7306	Near side			
W2690	350CAB7315	Far side			
W2690	350CAB7315	Near side		(Hit)	Likely lucky false call

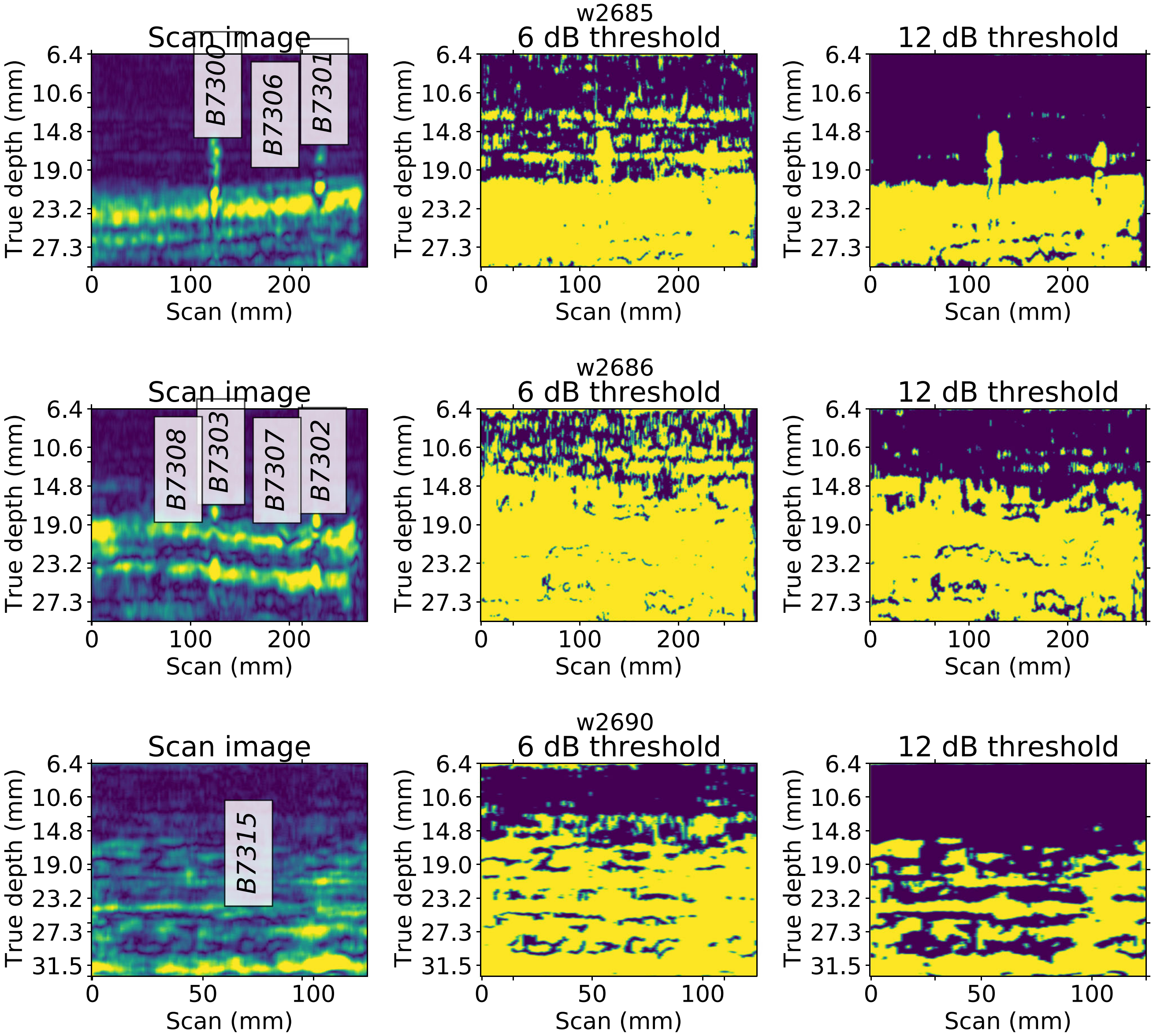


Fig. 6 Scan images of the validation welds from the near side. On the left the scan image and flaw sample numbers in their locations along the scan axis. On the middle indications above 6 dB noise threshold and on the right indications above 12 dB noise threshold

crack may have been due to lacking training data for the far side cases. However, the human inspector made more false calls in total than the ML model. On the other hand, the larger amount of false calls for the human inspector might related to the instructions given to the inspector. While the inspector was guided to avoid false calls, no minimum detection limit was set to determine the highest possible accuracy.

5 Conclusions

The following conclusions can be drawn from this study:

1. Rich multi-channel phased array ultrasonic data can be successfully used in automated flaw detection with modern machine learning network.
2. The machine learning networks can reach detection levels and $a_{90/95}$ values that are in line with what is expected from human inspectors.
3. Training data needs to represent the inspection case for reliable results.

Acknowledgements Welds were contributed by Suisto Engineering. UT data scanning was contributed by DEKRA. Data augmentation was contributed by Trueflaw. Their support is gratefully acknowledged.

Funding Open access funding provided by Technical Research Centre of Finland (VTT).

Data Availability The data sets generated and analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest. Virkkunen is associated with Trueflaw Ltd., who supplied the eFlaw augmentation.

References

1. Aggarwal, S.L.P.: Data augmentation in dermatology image recognition using machine learning. Skin Res. Technol. **25**(6), 815–820 (2019) https://doi.org/10.1111/srt.12726, https://onlinelibrary.wiley.com/doi/abs/10.1111/srt.12726
2. Ahmed, S., Reboud, C., Lhuillier, P.E., Calmon, P., Miorelli, R.: An adaptive sampling strategy for quasi real time crack characterization on eddy current testing signals. NDT E Int. **103**, 154–165,(2019) https://doi.org/10.1016/j.ndteint.2019.02.001, http://www.sciencedirect.com/science/article/pii/S0963869518304444
3. Ali, A.H., Balint, D., Temple, A., Leevers, P.: The reliability of defect sentencing in manual ultrasonic inspection. NDT E Int. **51**, 101–110 (2012). https://doi.org/10.1016/j.ndteint.2012.04.003
4. Bato, M.R., Hor, A., Rautureau, A., Bes, C.: Impact of human and environmental factors on the probability of detection during ndt control by eddy currents. Measurement **133**, 222–232 (2019) https://doi.org/10.1016/j.measurement.2018.10.008, http://www.sciencedirect.com/science/article/pii/S0263224118309242
5. Berke, M., Buechler, J.: Practical experiences in manual ultrasonic phased array inspection. In: WCNDT 2008, 17th World Conference on NDT—2008—Shanghai (China) (2008) https://www.ndt.net/article/wcndt2008/papers/110.pdf
6. Chen, F.C., Jahanshahi, M.R.: Nb-cnn: deep learning-based crack detection using convolutional neural network and Naive Bayes data fusion. IEEE Trans. Ind. Electron. **65**(5), 4392–4400 (2018)
7. Chen, F.C., Jahanshahi, M.R.: Real-time crack detection from nuclear inspection videos using fully convolutional network and parametric data fusion. IEEE Trans. Instrum. Meas. 1–1 (2020)
8. Chen, H., Lee, G.G.: Neural networks for ultrasonic nde signal classification using time-frequency analysis. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (1993)
9. Chowdhury, A., Kautz, E., Yener, B., Lewis, D.: Image driven machine learning methods for microstructure recognition. Comput. Materia. Sci. **123**, 176–187 (2016) https://doi.org/10.1016/j.commatsci.2016.05.034, http://www.sciencedirect.com/science/article/pii/S0927025616302695
10. Cochran, S.: Part 12. fundamentals of ultrasonic phased arrays. Insight—Non-Destructive Testing and Condition Monitoring 48 (2006) https://doi.org/10.1784/insi.2006.48.4.212
11. Cruz, F.C., Simas Filho, E.F., Albuquerque, M.C., Silva, I.C., Farias, C.T., Gouvea, L.L: Efficient feature selection for neural network based detection of flaws in steel welded joints using ultrasound testing. Ultrasonics **73**, 1–8 (2017) https://doi.org/10.1016/j.ultras.2016.08.017, https://www.ncbi.nlm.nih.gov/pubmed/27592203
12. Delannoy, L., Poirier, J., Hayes, P.: Manual phased array ultrasonic technic for pipe weld inspection in nuclear power plants. In: Proceedings of the 10th European Conference on Non-Destructive Testing (ECNDT 2010), NDT.net, The e-Journal of Nondestructive Testing (2010) https://www.ndt.net/article/ecndt2010/reports/1_03_55.pdf
13. DISSIMILAR Guidelines for generating array ultrasonic procedures for the inspection of dissimilar/austenitic welded components. TSBProject No: TP11/MFE/6/I/AA058J (2011)
14. Du, W., Shen, H., Fu, J., Zhang, G., He, Q.: Approaches for improvement of the x-ray image defect detection of automobile casting aluminum parts based on deep learning. NDT E Int. **107**, 102144 (2019)
15. Fei, C., Han, Z., Dong, J.: An ultrasonic flaw-classification system with wavelet-packet decomposition, a mutative scale chaotic genetic algorithm, and a support vector machine and its application to petroleum-transporting pipelines. Russ. J. Nondestr. Test. **42**(3), 190–197 (2006). https://doi.org/10.1134/s1061830906030077
16. Guan, X., He, J., Rasselkorde, E.M., Zhang, J., Abbasi, W.A., Zhou, S.K.: Probabilistic fatigue life prediction and structural reliability evaluation of turbine rotors integrating an automated ultrasonic inspection system. J. Nondestruct. Eval. **33**, 51–61 (2014) https://doi.org/10.1007/s10921-013-0202-z, http://www.sciencedirect.com/science/article/pii/S0893608015001446
17. Holmes, C., Drinkwater, B.W., Wilcox, P.D.: Post-processing of the full matrix of ultrasonic transmit-receive array data for non-destructive evaluation. NDT E Int. **38**(8), 701–711 (2005)
18. Hunter, A.J., Drinkwater, B.W., Wilcox, P.D.: The wavenumber algorithm for full-matrix imaging using an ultrasonic array. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **55**(11), 2450–2462 (2008)
19. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. (2015) CoRR abs/1502.03167, arXiv:1502.03167
20. Luo, H., Chen, Q., Lin, W.: Graphic augmented defect recognition for phased array ultrasonic testing on tubular tky joints. J. Nondestruct. Eval. **39** (2020). https://doi.org/10.1007/s10921-020-00698-x
21. Maier, H.P., Mletzko, U., Just, T., Neundorf, B.: Results of a round robin test on ndt methods for austenitic pipe welds in nuclear power plants. In: 2nd International Conference on NDE - Relation to Structural Integrity for Nuclear and Pressurised Components - 2000 - New Orleans (USA) (JRC-NDE 2000), NDT.net, The e-Journal of Nondestructive Testing, (2000) https://www.ndt.net/article/v05n08/neundorf/neundorf.htm
22. Masnata, A., Sunser, M.: Neural network classification of flaws detected by ultrasonic means. NDT E Int. **29**(2), 87–93 (1996)
23. Matz, V., Kreidl, M., Smid, R.: Classification of ultrasonic signals. Int. J. Mater. **27**, 145 (2006). https://doi.org/10.1504/IJMPT.2006.011267
24. Miorelli, R., Artusi, X., Reboud, C.: An efficient adaptive database sampling strategy with applications to eddy current signals. Simul. Model. Pract. Theory **80**, 75–88 (2018) https://doi.org/10.1016/j.simpat.2017.10.003, http://www.sciencedirect.com/science/article/pii/S1569190X17301430

25. Munir, N., Kim, H., Park, J., Song, S., Kang, S.: Convolutional neural network for ultrasonic weldment flaw classification in noisy conditions. Ultrasonics (2018)
26. Munir, N., Kim, H.J., Song, S.J., Kang, S.S.: Investigation of deep neural network with drop out for ultrasonic flaw classification in weldments. J. Mech. Sci. Technol. **32**(7), 3073–3080 (2018b)
27. Redmon, J., Farhadi, A.: Yolo9000: Better, faster, stronger. (2016) arXiv:1612.08242v1
28. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. (2018) arXiv:1804.02767v1
29. Sambath, S., Nagaraj, P., Selvakumar, N.: Automatic defect classification in ultrasonic ndt using artificial intelligence. J. Nondestr. Eval. **30**(1), 20–28 (2010). https://doi.org/10.1007/s10921-010-0086-0
30. Schmerr Jr., L.W.: Fundamentals of Ultrasonic Phased Arrays. Springer, New York (2015)
31. Shukla, K., Leoni, P.C.D., Blackshire, J., Sparkman, D., Karniadakis, G.E.: Physics-informed neural network for ultrasound nondestructive quantification of surface breaking cracks. J. Nondestr. Evalua. **39** (2020). https://doi.org/10.1007/s10921-020-00705-1
32. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. (2014) eprint arXiv:1409.1556v4
33. Spencer, F., Schurman, D.: (US) FAATC, Laboratories SN Reliability Assessment at Airline Inspection Facilities, Volume III: Results of an eddy current inspection reliability experiment. nid. 3, Federal Aviation Administration Technical Center, (1995) https://apps.dtic.mil/dtic/tr/fulltext/u2/a295081.pdf
34. Svahn, P.H., Virkkunen, I., Snögren, D., Zettervall, T.: The use of virtual flaws to increase flexibility of qualification. In: 12th European Conference on Non-Destructive Testing (ECNDT 2018), NDT.net, The e-Journal of Nondestructive Testing (2018)
35. Tang, S., Chen, Z.: Scale-space data augmentation for deep transfer learning of crack damage from small sized datasets. J. Nondestr. Eval. **39** (2020). https://doi.org/10.1007/s10921-020-00715-z
36. Virkkunen, I.: Virtual cracks and the future of inspection reliability. In: Kärnteknikdagarna Nordic Symposium on Nuclear Technology (2017)
37. Virkkunen, I., Miettinen, K., Packalén, T.: Virtual flaws for nde training and qualification. In: 11th European Conference on Non-Destructive Testing (ECNDT 2014), NDT.net, The e-Journal of Nondestructive Testing (2014)
38. Virkkunen, I., Rönneteg, U., Grybäck, T., Emilsson, G., Miettinen, K.: Feasibility study of using eflaws on qualification of nuclear spent fuel disposal canister inspection. In: 12th International Conference on NDE in Relation to Structural Integrity for Nuclear and Pressurized Components (JRC-NDE 2016), NDT.net, The e-Journal of Nondestructive Testing (2016)
39. Virkkunen, I., Koskinen, T., Jessen-Juhler, O.: Virtual round robin – a new opportunity to study ndt reliability. Nucl. Eng. Des. **380**, 111297, (2021) https://doi.org/10.1016/j.nucengdes.2021.111297, https://www.sciencedirect.com/science/article/pii/S0029549321002491
40. Virkkunen, I., Koskinen, T., Jessen-Juhler, O., Rinta-Aho, J.: Augmented ultrasonic data for machine learning. J. Nondestr. Eval. **40**(4) (2021) https://doi.org/10.1007/s10921-020-00739-5
41. Yi, W., Is, Yun: The defect detection and non-destructive evaluation in weld zone of austenitic stainless steel 304 using neural network-ultrasonic wave. KSMME Int. J. **12**(6), 1150–1161 (1998)
42. Zhu, P., Cheng, Y., Banerjee, P., Tamburrino, A., Deng, Y.: A novel machine learning model for eddy current testing with uncertainty. NDT E Int. **101**, 104–112 (2019). https://doi.org/10.1016/j.ndteint.2018.09.010

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.