



The Effect of Different Flaw Data to Machine Learning Powered Ultrasonic Inspection

Tuomas Koskinen^{1,2} · Iikka Virkkunen² · Oskar Siljama² · Oskari Jessen-Juhler¹

Received: 3 August 2020 / Accepted: 30 January 2021 / Published online: 18 February 2021
© The Author(s) 2021

Abstract

Previous research (Li et al., Understanding the disharmony between dropout and batch normalization by variance shift. CoRR abs/1801.05134 (2018). <http://arxiv.org/abs/1801.05134> arXiv:1801.05134) has shown the plausibility of using a modern deep convolutional neural network to detect flaws from phased-array ultrasonic data. This brings the repeatability and effectiveness of automated systems to complex ultrasonic signal evaluation, previously done exclusively by human inspectors. The major breakthrough was to use virtual flaws to generate ample flaw data for the teaching of the algorithm. This enabled the use of raw ultrasonic scan data for detection and to leverage some of the approaches used in machine learning for image recognition. Unlike traditional image recognition, training data for ultrasonic inspection is scarce. While virtual flaws allow us to broaden the data considerably, original flaws with proper flaw-size distribution are still required. This is of course the same for training human inspectors. The training of human inspectors is usually done with easily manufacturable flaws such as side-drilled holes and EDM notches. While the difference between these easily manufactured artificial flaws and real flaws is obvious, human inspectors still manage to train with them and perform well in real inspection scenarios. In the present work, we use a modern, deep convolutional neural network to detect flaws from phased-array ultrasonic data and compare the results achieved from different training data obtained from various artificial flaws. The model demonstrated good generalization capability toward flaw sizes larger than the original training data, and the effect of the minimum flaw size in the data set affects the $a_{90/95}$ value. This work also demonstrates how different artificial flaws, solidification cracks, EDM notch and simple simulated flaws generalize differently.

Keywords NDT · Ultrasonic testing · Machine Learning · Image classification

1 Introduction

Ultrasonic inspectors are commonly trained using simple artificial flaws, such as EDM notches and side-drilled holes. These two types offer a quick and cost-effective way of demonstrating where the flaw indication should appear, but their signal shape differs from a real service-induced crack, like a mechanical or thermal fatigue crack. Inspectors can use reasoning to estimate real reflectors based on these simplified signals. However, more than simple artificial flaws are

usually required for qualification of a technique, for example in nuclear power plants [7], to confirm performance in a representative setting.

Difficulties arise when the inspection material is noisy and the inspector needs to use expert judgement to distinguish flaws from structural noise. For example, an EDM notch might be found much more easily than a thermal fatigue crack in a dissimilar metal weld (DMW) inspection. While their signals can be distinguished from each other, a human inspector is not only looking for a specific reflector or thermal fatigue crack but also for an explanation for any unusual reflector. Therefore, while training and conducting the inspections, the inspector focuses on learning and detecting where the flaw indications may appear and how they stand out compared to the surrounding noise. A human inspector can intuitively ignore possible artefacts in the artificial flaws and still successfully find real flaws in the inspection data.

✉ Tuomas Koskinen
tuomas.koskinen@vtt.fi
Iikka Virkkunen
iikka.virkkunen@aalto.fi

¹ VTT Technical Research Centre of Finland Ltd, VTT, P.O. Box 1000, 02044 Espoo, Finland

² Aalto University, P.O. Box 11000, 00076 Aalto, Finland

For machine learning, the task is much more difficult. Due to the training process, the model can learn any and all features related to the training data; thus, the teaching data set determines the boundaries of the capability of the algorithm. This learning method is useful when the task is to determine specific features from images with high accuracy and the training data are largely available. For ultrasonic inspection, this is a problem since training data are not readily available and the detection probability for the algorithm needs to be high, while still avoiding false calls. The model may learn incidental features of the training data, i.e. it may overfit to the training flaws and fail to generalize to unseen flaw indications. Conversely, underfitting may cause an excessive false call rate. Therefore, this paper aims to study how training data from different sources can be used to train ML algorithms to detect other flaw types and how the minimum flaw size in the training set affects the $a_{90/95}$ value.

1.1 Effect of Different Kinds of Flaws and Artificial Flaws

The flaw response for ultrasonic testing is highly related to the kind of reflector from which the sound waves are reflected back to the transducer. The characteristics that primarily affect the flaw response are the location and orientation of the crack, size of the crack, opening of the crack through the whole path and at the crack tip, fracture surface roughness and filling of the crack with a substance. An in-depth study of the flaw responses and crack characteristics has already been conducted by [9,10]. For the most representative flaw response signal, it is reasonable to assume that these characteristics should be met in order to achieve the best possible training data for machine learning as well. These characteristics are the main reason real cracks are preferred over EDM notches and side-drilled holes when conducting actual performance demonstrations for human inspectors. In addition, it is assumed that the larger the crack, the easier it is to detect. This should apply for the machine learning model as well. As the larger cracks are more critical, these types of cracks should be reliably found.

1.2 Teaching and Generalizing the Machine Learning Model

Since humans can use their theoretical reasoning and target their focus on the relevant part of the data, it is possible (to a certain extent) to use simple flaws to teach and train human inspectors to find real flaws in inspection cases. Machine learning models lack this theoretical reasoning and imagination, and the training data must explicitly provide the variation that the models need to learn.

The training data itself has a strong influence on training and generalizing the model. First of all, it is imperative to

have enough flaw data for teaching. Secondly, the labelling of the data needs special attention to non-destructive testing (NDT). Labelling small flaws that are indistinguishable from the noise may cause the model to overfit on noise features and/or result in an excessive false call rate. Lastly, the models can converge in training, even in the absence of generalizable features in the training data, as demonstrated by [26].

Overfitting can be mitigated by several approaches. The obvious first choice is to increase the amount of training and validation data. This ample data amount is seldom available for ultrasonic testing. The second choice for decreasing overfitting is data augmentation when teaching data are scarce. Traditional data augmentation, where the image is rotated, reflected, scaled, cropped or translated, are common practices to artificially increase the amount of available data [3,6]. These methods have been used successfully in NDT and ultrasonic inspection by [25]. For a weld scan, however, rotation of the flaw might be out of the question as cracks can form in a certain place and certain orientation for in-service inspections. Data augmentation through virtual flaws presented in [23] has shown great promise as it allows scaling the flaws to represent smaller flaws and changing the location along the weld, allowing a larger variety of backgrounds for the flaw to reside in.

In general, NDT data can be considered simple, thus there exist options for generating data other than virtual flaws. Alternative approaches for generating training data sets have been used in eddy current testing by [15] to generate an ample amount of data with an adaptive generation technique known as Output Space Filling (OSF) with an efficient computation time. Reference [1] used a similar approach by adding the Partial Least Squares (PLS) feature extraction to OSF and trained several machine learning models with this generated data. Reference [1] stated that this data generation method might be feasible for ultrasonic and thermographic testing.

Further generalization can be obtained by tuning the hyperparameters. Batch size, for example, has a strong influence on learning. Reference [14] showed that generally, the best generalization performance is achieved with a batch size of 2 to 32 and up to 64 with a batch normalization layer. The downside of using small batch sizes is that it slows down the teaching of the model. Thus, large batch sizes are preferred.

Instead of further modifying or augmenting the teaching data or decreasing the performance of the model, there are also possibilities to affect the training of the model as well. Dropout is one of the most common approaches. Dropout works by zeroing out a certain amount of the layer's output values at random during training. The number of values dropped out is determined by the dropout rate, which is usually between 10 and 50% of the layer's output values. Essentially, this means that random variation is introduced to the output, and less significant features that are only present for the training data are valued less or cancelled out from the

final model, thus leading to a more generalized representation of the task. During testing, the output values are allowed to work fully, but scaled down with the dropout rate to compensate all working output values [6,24]. Dropout has been used in ultrasonic inspection by [16] with successful results as the performance increase was significant compared to the neural network without the dropout for the A-signal classification.

As overfitting is one of the major problems in teaching modern deep learning models, the difference between human image recognition and machine image recognition needs to be understood as well. Unlike a human inspector, a machine learning model does not know the actual concept of its task, since that is determined through the teaching data. Even though the data might look good enough for teaching humans and estimating the probability of detection (POD) curves [12,21], the data might contain artefacts from the artificial flaw manufacturing process or poorly designed virtual flaw generation. This kind of teaching, i.e., with poor data, is demonstrated in [17], where distinguishing between wolves and huskies was based on the feature that wolves had snow in the background in the training data set. The effect of the poor data is the same for NDT. When the teaching data would have some feature such as an artefact from implantation, the flaw detection in the model might focus on the implantation rather than the actual flaw characteristics. Due to these reasons, it is crucial for the NDT model to be actually tested with flaws where these kinds of artefacts do not exist or to map which features affect the decision the most with methods such as or similar to grad-CAM by [18] and LIME by [17].

Generalization of the model can be increased through adding the batch normalization layer introduced in [8]. Batch normalization drives to remove the covariate shift from the internal activations within the network. This has the effect of faster learning rates and increased accuracy. As batch normalization works to generalize the model, it decreases the need for a dropout layer in some cases. In fact, the performance of the model might decrease drastically if a dropout and a batch normalization layers are used together. Reference [13] recommend the use of a dropout layer after all batch normalization layers on large data sets. On the other hand, Reference [5] reported a decrease in accuracy when both layers were used together. In general, it is recommended to use a batch normalization in the models first and then carefully observe the effect of an added dropout layer for the best possible result.

Therefore, the main problem is teaching a model with too little real flaw data, while still keeping generalization to real flaws that the model has never seen before and still maintaining at least human-level performance. As the flaw data is scarce in NDT, virtual flaws present a way to mitigate the problem. However, the more diverse the data, the better, even with the virtual flaws. Hence, simulating the flaw responses for training might be plausible to broaden the data efficiently.

Simulated flaws have been previously used together with virtual flaw augmentation to calculate POD by [11]. While humans did not detect the difference between simulated and real flaw responses, the research showed that simulated flaws were slightly easier to detect. Thus, it might be assumed that an ML model could be able to adequately generalize to real flaws.

2 Materials and Methods

Inspection data was gathered from scanning a DMW mock-up and generating flaw responses from CIVA simulations. The location of the flaws was the same for all flaw types, on the edge of the buffer zone, 7 to 10 mm to the carbon steel from the weld center. The scanned flaws were augmented with Trueflaw's eFlaw [21] software, and data sets for machine learning purposes were created.

2.1 Scanned Samples

For initial inspection data, a DMW pipe mock-up from Swedish Qualification Centre AB (SQC) was used. The specimen was 32 mm thick with an outer diameter of 898 mm. The specimen had implanted flaws and an EDM notch as defects. The original sample consisted of two "small" solidification flaws 2 mm and 3 mm in size. Two large solidification flaws, of which 17 mm was tilted toward the carbon steel side and 26 mm was straight oriented. There were two 6 mm sized flaws, an EDM notch and a solidification flaw. In total, six different flaws were available for training. In addition, the sample consisted of three axial solidification flaws with heights of 6, 17 and 26 mm and one axial EDM notch with a height of 6 mm. The flaw scanning was optimized for circumferential flaws; thus, the axial flaw indications were removed from the teaching and testing data sets with the eFlaw process.

The inspection procedure was an optimized version of Zetec Inc.'s procedure C3467 Zetec OmniscanPA 03 Rev A. The inspection equipment that was used was Dynaray Lite with two Imasonic 1.5 MHz 32 element phased array probes in a wedge with a 7° roof angle set-up for TRL acquisition. The coupling was applied through a feed water system. In order to minimize data, only one scan line was utilized, with a 60° angle. The focal law was focused on the inner surface of the pipe and the probe positioned such that the best amplitude response from the flaws was achieved. Data recording was done at a 16-bit depth for best possible data quality. The schematic of the inspection procedure can be seen in Fig. 1. The scanned flaws were augmented with eFlaw software by scaling down the recorded amplitude, thus representing a wider size range of flaws similarly as in [22].

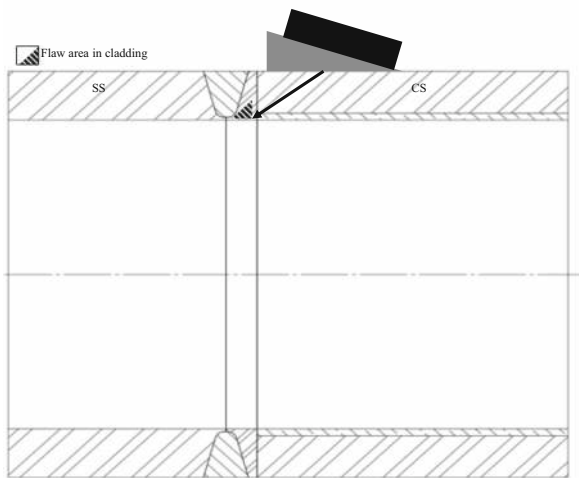


Fig. 1 Diagram of the inspection setup. TRL probe was situated on the carbon steel side of the test mock-up and focused on inner diameter of the pipe. The original flaws were situated on the cladding marked with black stripes on the image. Due to export control restrictions, the exact details of the test block's dimensions, materials and weld cannot be made public

2.2 Simulation Set-up

The same set-up was created in CIVA2019 simulation software with different-sized notches. For signal generation, the Hanning type was used, and for flaw response calculation, the Kirchhoff and GTD model was used as per CIVA guidance [4,20] in a similar simulation case, which is optimal for simulating reflection and diffraction echoes from crack like flaws. Unlike the paper by [20], the weld was modelled with orthotropic anisotropy. The buttering layer of the DMW was modelled with a polycrystalline cubic structure, with an average grain size of $1.5 \mu\text{m}$ to represent the simple simulation case. In order to reduce calculation time, only the flaw and the immediate surroundings of the flaw were simulated. The resolution of the simulation was aimed to be the same as for the scanned samples' 2 mm scan step and 103.2 mm sound path.

In total, six different-sized notches were simulated; at heights of 1 through 6 mm, the width of the flaw was three times the height. Just as with the scanned plate samples, the flaw responses were extracted from the simulation data and implemented in the pipe mock-up scan through the eFlaw software and more flaws were generated by scaling down the recorded amplitude from the simulated flaws to a total of 10,000 simulated flaws generated by the eFlaw augmentation.

Figure 2a demonstrates the original simulated B-scan image from CIVA and Fig. 2b the pre-processed B-scan image for a better comparison. The raw simulated signal was used when implanting the flaw onto the scanned image with eFlaw and pre-processed for model training. Figure 2c shows

the pre-processed simulated flaw image show to the model and Fig. 2d pre-processed scanned EDM notch implanted with eFlaw. The width of the simulated flaw matches well with the scanned one. Along the sound path, the simulated flaw is slightly longer; and after post-processing, the simulated flaw looks denser than the scanned EDM notch.

As Fig. 2 demonstrates, the scanned 6 mm EDM notch and the simulated 6 mm EDM notch look different. This is because the aim was to use a simple simulation setup in CIVA to set a base-line for teaching data. The size of the flaw is accurate, and the implanted signal is plausible for human eye as well due to the accurate modelling of the flaw and model geometry. However, a closer simulation could be achieved with increased accuracy in the material and anisotropy parameters of the DMW as well as increased detail in the simulated signal representing the used probe more accurately as only the frequency was matched to represent the scanned signal.

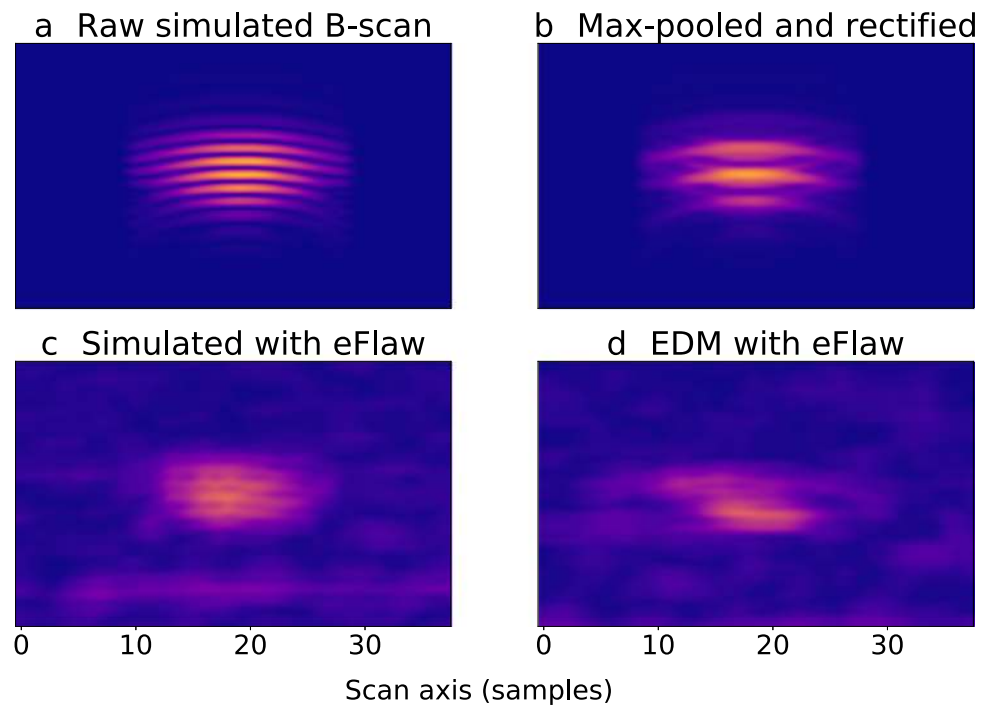
2.3 Training Data and Used Data Augmentation

Reference [23] used only thermal fatigue flaws as scan input; thus, it is proven that it is viable to use thermal fatigue flaws as teaching material to find thermal fatigue flaws. For this paper, we generated several different teaching data sets, where certain flaw types were only shown during testing to investigate how well the model detects the completely new flaw type.

In order to generate sufficient training data from the six different scanned flaws and six different simulated flaws, eFlaw software was used to augment the flaw locations and sizes within the training and testing data. These virtual flaws have been previously used successfully in training humans and evaluating POD by [12,19,21]. The indications of the six different scanned flaws and simulated flaws were scaled down to represent smaller sizes up to 40% of the original indication. This allowed the generation of 7000 different variations for the scanned data to be used as training, validation and testing data with roughly 50% containing flaws and 50% without flaws. In addition, great care was taken to prevent the model from learning the virtual flaw introduction process by copying and replacing the unflawed data as well within the set. A total of 10,000 images were created from the simulated flaws with the same method as for the scanned flaws.

The raw RF signal was pre-processed by fully rectifying the signal to an absolute positive value. The scan data was processed for more efficient teaching purposes, thus the sound path was narrowed down to 2000 samples to represent the inner diameter of the mock-up where the flaws were located. The flaw image contained 480 scan steps in total. The B-scan dimension of 480×2000 proved to be too slow to handle, as the whole data set could not fit into the GPU memory at the same time. In order to reduce the data set size, without losing information from the sound path, the original B-scan

Fig. 2 Comparison of the simulated 6 mm EDM notch signal and scanned 6 mm EDM notch implanted through the eFlaw software. **a** Raw simulated RF signal with sound path of 2000 samples, **b** raw simulated signal post processed with max-pooling and rectified to absolute positive value, **c** simulated flaw implanted to the weld b-scan with eFlaw and **d** scanned 6 mm EDM notch implanted to the weld B-scan with eFlaw for comparison. The simulated flaw seems to be slightly longer and denser along the sound path and more symmetric than the reference scanned EDM notch



was pre-processed by max-pooling the sound path with $\frac{1}{4}\lambda$. This provided original data in the size of 480×118 . To further optimize the image for machine learning, the image was normalized according to [3]. The image was reduced by the mean value and divided by the standard deviation. If the image was labelled as flawed, one flaw would be introduced in the image through the eFlaw process at a random location along the weld. Since only one weld was scanned as the background canvas, it later showed that the model learned the weld pattern when it was shown the whole 480-sample-wide weld in a single training image. This led the model to overfit on the weld rather than detecting the actual flaw indications. This is clearly the wrong target as this would work only if the actual inspected weld would provide an exactly identical weld image as that recorded from the mock-up, which is impossible. This was mitigated by cropping the image area in half. Once the image size was 240 samples wide, it allowed the generation of images at multiple locations along the weld and maintaining a generalization on the clean weld, as the background kept changing. This meant that the model was shown a “new” clean weld with no flaws as for the flawed samples as well; thus, the initial image data was 240×118 samples in size.

For a proper comparison to the previously mentioned VRR data, the model was adjusted to handle 48×118 sized images to determine the proper location from the data. As the image was smaller than the original teaching data, the sound path was further cropped to 112 samples. This allowed randomly moving the crop window along the sound path for 6 samples, increasing the different backgrounds for training data. The

variability of the images was further increased by a similar data augmentation used by [25]. The image was randomly flipped from left to right during training using the built in function from the Tensorflow package, but not rotated or scaled. To further validate that the taught model would not overfit, the images with no flaws were shown only 90% of the weld area. During testing the model would see the whole weld.

The model was trained with the following flaw type combinations from (a) through (f), shown in Table 1 for solidification cracks and an EDM notch. The model taught with the simulated flaws was run with two different types of combinations, (i) and (j) in Table 2. The tables show the amount of flaws available for training. The total number of images is doubled when the images without flaws is added to the data set. 20% of this said data set would be selected as the validation set. In addition, the model was taught with only 6 mm solidification crack (g) consisting of 558 flaw images and only a 6 mm EDM notch (h) consisting of 599 flaw images not included in the tables.

2.4 Used ML Architecture

A more refined deep neural network model was constructed based on [23]. To further enhance the accuracy, the dimensions of the latter convolutional layers and the dense layer were increased, and the max pooling layer with the batch normalization layer was added after each convolutional layer for increased generalization and overfitting reduction. The model architecture can be seen in Fig. 3. The optimized net-

Table 1 Flaw sets for training and flaw images

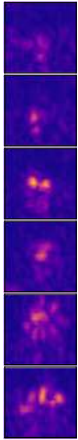
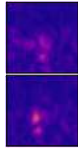
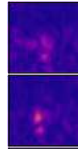



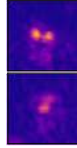

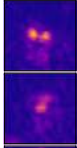





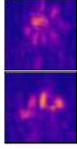




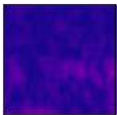
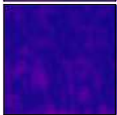
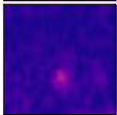
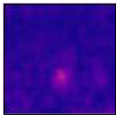
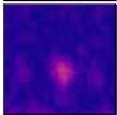
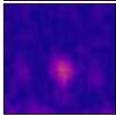
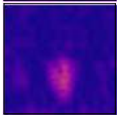
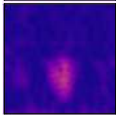
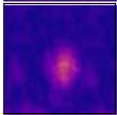
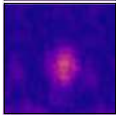
Size \ Set	a) All flaws	b) Small	c) Medium	d) Large	e) No larges	f) No smalls
Flaws total:	3442	1090	1157	1195	2247	2352
2 mm						
3 mm						
6 mm						
6 mm (EDM)						
17 mm						
26 mm						

Table 2 Simulated flaw sets for training and flaw images

Size \ Set	i) All	j) No smalls
Flaws total:	4982	3360
1 mm		
2 mm		
3 mm		
4 mm		
5 mm		
6 mm		

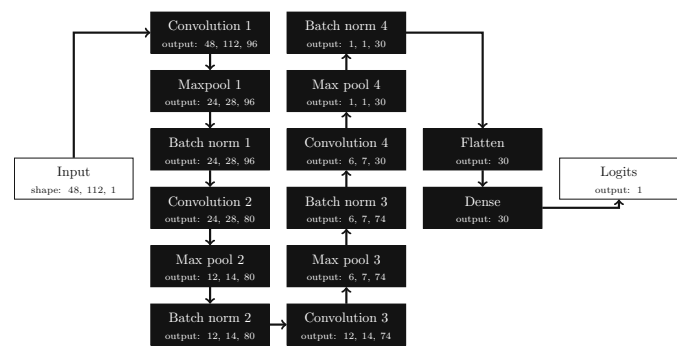
work structure was a result of trial and error by adjusting the dimensions of each layer. Dropout was left out of the model as batch normalization proved to be sufficient and using both dropout and batch normalization together seemed to yield variability in the results. The model was taught with the training data variations presented in Sect. 2.3.

2.5 Performance Evaluation

POD and false calls were used to measure the performance of the model. The POD curve was a hit/miss POD calculated according to regular standard MIL-HDBK-1823a [2]. POD is a valid way to measure the performance of the model, since it is used for evaluating the performance of humans as well. In addition, this enables comparison between the model result and human VRR data, since they were evaluated with the same data and standard. If the model were overly sensitive, it would show as false calls in the evaluation or if the model would overfit and constantly miss flaw types never seen before, this would easily be seen in the POD curve as erratic behaviour.

The performance evaluation was divided into two data sets from the virtual flaws described in Sect. 2.3. The first test data set would contain 4700 to 7000 samples, depending on which flaw set from Table 1 was used for training. Only the flaw types that were not used in training would be shown to the model to evaluate the generalization capability. The second data set would contain 1000 samples with all the available flaw types. Even though the same flaw types used in training are expected to be more easily found, they do not have an effect on finding the flaw types never shown to the model in training and are in the data set only to avoid flaw size gaps in the POD evaluation.

Fig. 3 Used optimized network structure



After testing with the data set, which contained images from random locations of the DMW with or without a flaw, the model was shown the same ultrasonic weld image as a human would see in a traditional inspection. This was done by dividing the whole weld image to 48-sample-wide images with the location coordinate as metadata. The images would be shown to the model, and the model would evaluate whether or not the location contains a flaw. In case of a hit, the image centre line would be highlighted in green on the weld image.

3 Results

The results have been divided into two sections: testing the generalization capability of the model; and comparison to human performance with similar ultrasonic data.

3.1 Testing the Generalization

For POD calculations, the data was adjusted similarly as in [22] when there were no missed flaws; 0 to 0.2 mm sized misses were added to the calculation. When the data faced zero separation, a missed flaw was added with a size 0.1 mm larger than the smallest flaw found. These adjustments needed to be done for the POD calculation to converge some of the results. However, this has little effect on the final POD.

The predictions for the differently trained models have been plotted in Fig. 4. The model was trained with the flaw combinations (a)–(j) described in Sect. 2.3 and in Tables 1 and 2. The tested flaws have not been shown to the said model before. This enables testing how well the model is capable of generalization when trained with different flaws and tested with completely different flaws. For POD hit/miss evaluation, all the indications scoring higher than 50% was considered as hits, and false calls when no flaw was in the data. If the prediction was less than 50% and flaw existed in the data, it was marked as a miss. The POD was tested with a data set containing all the flaw types and 1000 samples. The PODs for different models can be found in Fig. 5, except for the simulated flaws, as the performance was so unreliable due to false calls that it was not comparable to other models.

For cases (a), (b) and (e) in Fig. 5, adding zeros from size 0 to 0.2 mm and zero separation management have been used due to low or no misses. (a) and (e) gave no POD before the adjustment and (b) changed to a more conservative $a_{90/95}$ value of 1.05 mm from 0.75 mm.

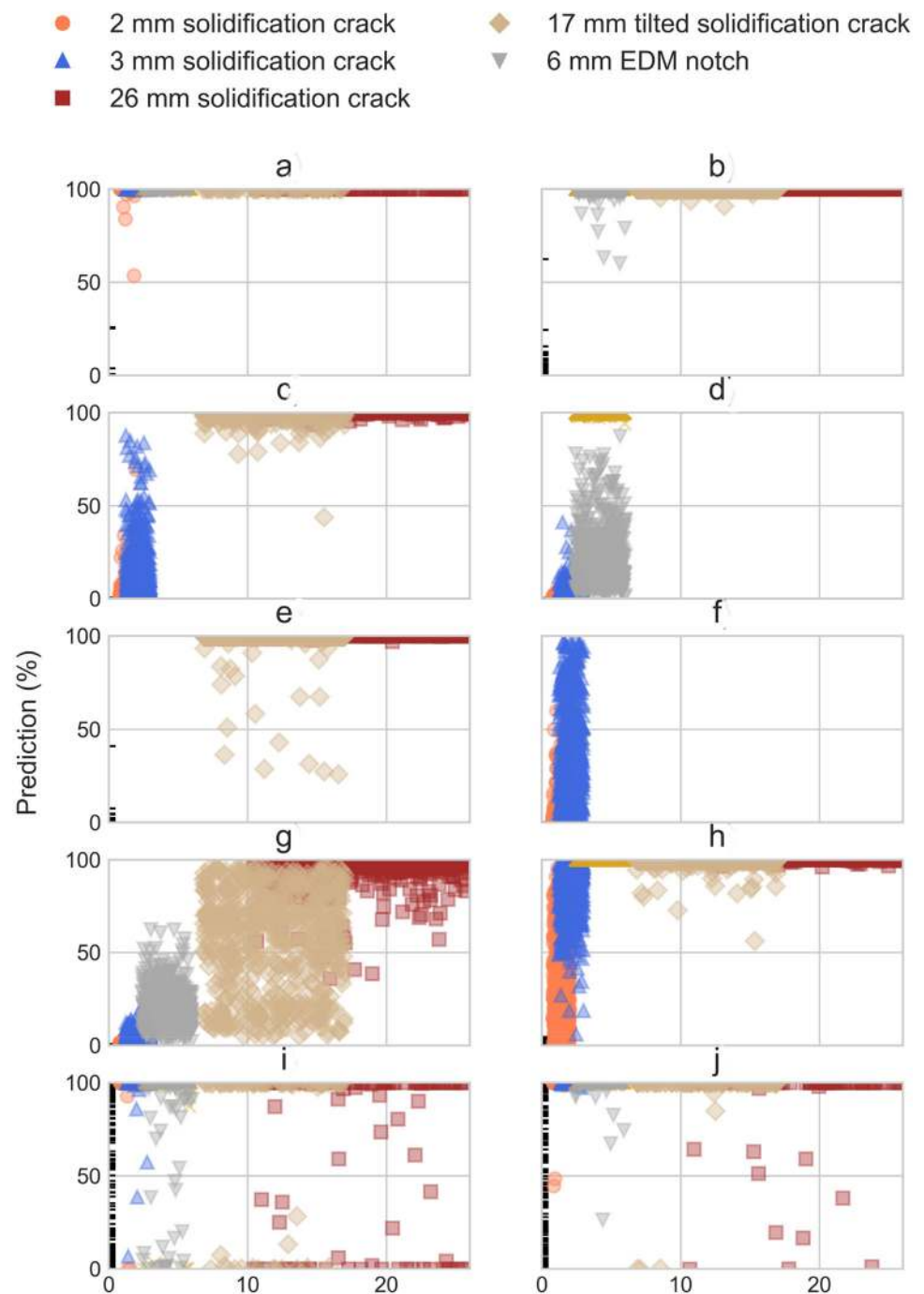
As expected, the model trained with all the available flaw types provided perfect results with no missed flaws and zero false calls. This test was done to set the benchmark for other training data sets.

When training with only the smallest flaws, the model generalizes well on the larger flaws. Predictions for the flaws can be seen in Fig. 4b. There is a slight deviation for the predictions for the EDM notch and 17 mm solidification crack, which was slightly tilted compared to the 6 mm and 26 mm solidification cracks, both of which yielded almost perfect predictions but also had with three reported false calls. The POD for the model trained with only small flaws can be seen in Fig. 5b for which no misses on the 17 mm flaws were reported while few of the smaller flaws were missed. The POD is the same as for the case with all available flaws (a), since the smallest flaw available for training is the same.

When the model was trained with the 6 mm solidification crack and the 6 mm EDM notch, the model managed to generalize well on the larger flaws, while generalization towards the smaller solidification cracks was not impressive. As seen in Fig. 4c the model missed one 17 mm solidification crack for the larger testing set but found all with the smaller test set size for POD. This is well within the $a_{90/95}$ limit, as the larger test set contained over 500 samples of virtually augmented 17 mm solidification cracks, which were not shown to the model during the training of the (c) set. The augmentation for this flaw group was from 2.4 to 6 mm, and the $a_{90/95}$ value was 3.45 mm, which can be seen in Fig. 5c.

The results for the model trained with only large flaws yielded similar results in Fig. 4d and the POD in Fig. 5d. Through the augmentation process, the smallest flaw for training was 6.8 mm. The model did poorly in finding smaller flaws, with the exception of the 6 mm solidification cracks, for which every flaw was found. This also has a decreasing effect on the POD and $a_{90/95}$ values.

Fig. 4 Predictions vs. flaw size when testing with unseen flaws, with exception to **a**. Threshold for detection was set to 50% **a** all flaws used for training. All flaws found with no false calls. **b** Only small flaws, 2 and 3 mm used for training. All flaws found, 3 false calls. **c** Only medium flaws, 6 mm solidification crack and EDM notch used for training. One miss on 17 mm flaw, no false calls, poor performance on smaller flaws. **d** Only large flaws, 17 and 26 mm used for training. Only 6 mm solidification cracks found reliably, no false calls. **e** Large flaws, 17 mm and 26 mm removed for training. Some of the 17 mm flaws missed, no false calls. **f** Small flaws, 2 and 3 mm removed for training. Smaller flaws are not found with high consistency, no false calls. **g** Only 6 mm solidification crack used for training. Only largest flaws are found reliably, consistent misses on 6 mm EDM notch, no false calls. **h** Only 6 mm EDM notch used for training. Generalizes well on larger flaws and 6 mm solidification crack, missing constantly smaller flaws, no false calls. **i** Trained with all simulated flaws. 131 False calls, missing constantly every flaw type. **j** Small simulated flaws removed. 78 Calls, misses from every flaw type, slightly better performance compared to **i**

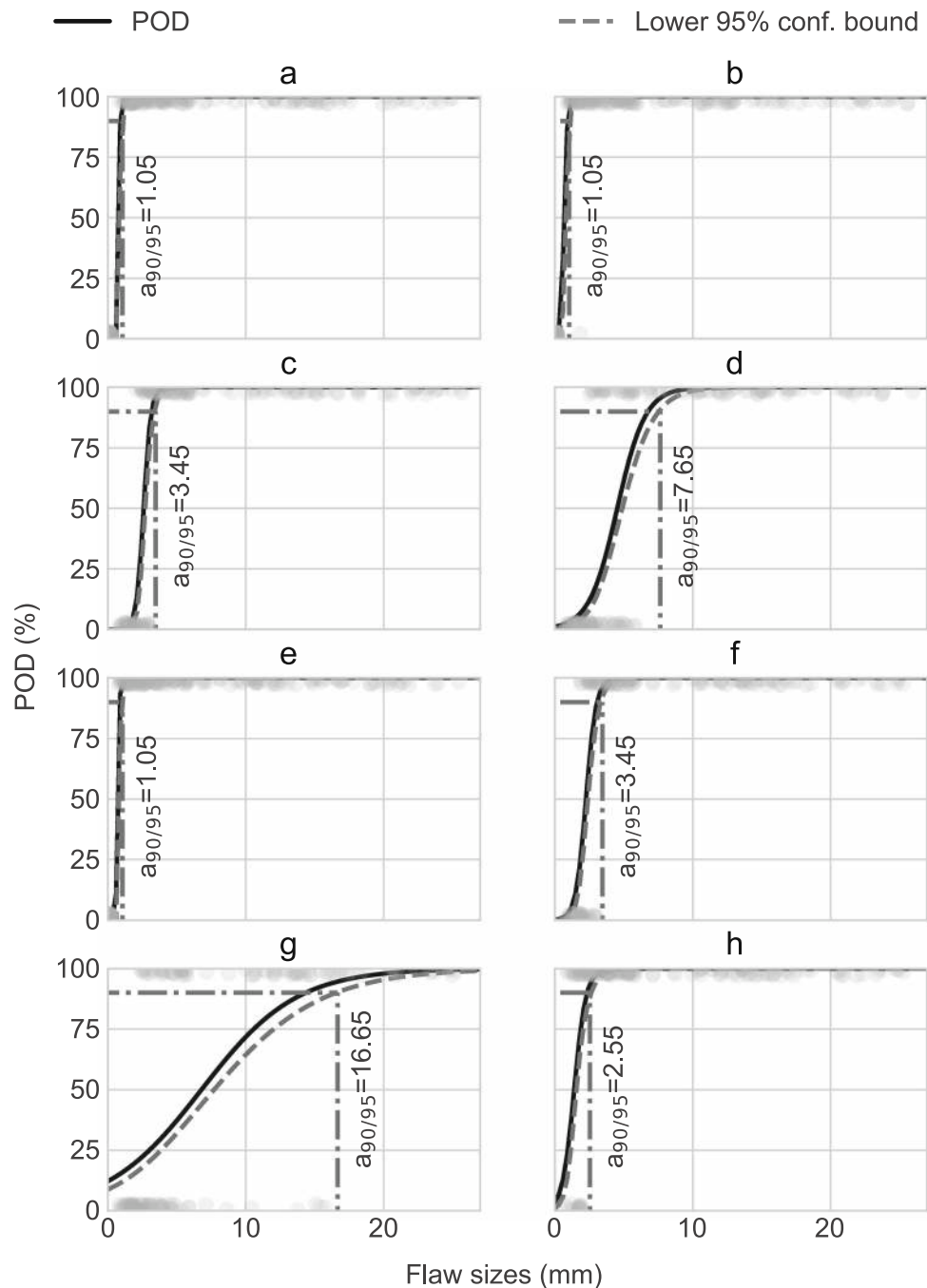


When training without the large flaws, the model generalizes well with the larger flaws. Predictions for the flaws can be seen in Fig. 4e. There is a slight deviation for the predictions for the 17 mm solidification crack, as similarly seen when training only with the medium sized flaw in Fig. 4c, compared to 26 mm solidification cracks, which yielded perfect predictions. Again, the testing set for never before seen flaws was significantly larger than the testing set for POD measurement, which contained all the flaw types. Thus, missing 6 of

the 17 mm flaws in a testing set containing over 500 flaws is within statistical limits. The POD for the model trained without the large flaws can be seen in Fig. 5e where no misses on 17 mm flaws were reported, which meant that the model made no misses or false calls.

When the small flaws were excluded from training, the result was the same as when training with only the medium sized flaws in Fig. 4c. Predictions can be seen in Fig. 4f and the POD in Fig. 5f.

Fig. 5 POD when testing with all flaw types. **a** All flaws used for training. All flaws found with no false calls. 0 to 0.2 mm sized misses added for POD convergence. **b** Only small flaws, 2 and 3 mm used for training. Two 2 mm sized flaws missed with no false calls. 0 to 0.2 mm misses added for more conservative POD. **c** Only medium flaws, 6 mm solidification crack and EDM notch used for training. Smallest flaw size in training set was 2.4 mm and $a_{90/95}$ was 3.45 mm. **d** Only large flaws, 17 and 26 mm used for training. As the 6 mm solidification cracks are found reliably, it improves the POD whereas the 6 mm EDM notches are mostly missed. **e** Large flaws, 17 mm and 26 mm removed from the training set. No false calls and 0 to 0.2 mm sized misses added for convergence. **f** Small flaws, 2 and 3 mm removed from training set. Smaller flaws are not found with high consistency, same $a_{90/95}$ result as in **c** since smallest flaw size was the same. **g** Only 6 mm solidification crack used for training. Only largest flaws are found reliably, consistent misses on 6 mm EDM notch, the worst $a_{90/95}$. **h** Only 6 mm EDM notch used for training. Generalizes well on larger flaws and 6 mm solidification crack, missing smaller flaws, but performs better than **c** and **f** with $a_{90/95}$ of 2.55



When using only one flaw type, the training set tends to get dangerously small. When the model was trained with only a 6 mm solidification crack, the result deteriorates considerably. All the smaller cracks are missed, EDM notches are barely detected and the model struggles to detect the larger 17 mm and 26 mm cracks without false calls. The predictions can be seen in Fig. 4g and the POD in Fig. 5g. However, the model trained with only a 6 mm EDM notch proved to perform well compared to the same-sized solidification crack. The predictions of the model trained with only a 6 mm EDM

notch can be seen in Fig. 4h and the POD in Fig. 5h. There is a slight deviation for the 17 mm solidification crack, but 6 mm and 26 mm are found consistently. The majority of the 3 mm solidification cracks are found, while the 2 mm solidification crack tends to go unnoticed, which has an improving effect on the POD. When training with only the EDM notch, the model achieves the best $a_{90/95}$ value when there are no small flaws included in the training set.

The results when training with all the simulated flaws can be seen in Fig. 4i and results for the training with simulated

flaws without the smaller flaws in Fig. 4j. A generalization to real defects proved to be not possible for the simple simulated flaw. Even though the flaw sizes ranged from 1 to 6 mm, there were no indications of improvement, as false calls were an issue when tested with real flaws. The most concerning observation was that, while the model was capable of generalization to larger flaws from smaller flaws, this was not the case for the simulated flaws. The model kept constantly missing some of the largest flaws for both training data sets. The only difference was that when the smallest flaws were removed from the data set, performance increased by decreasing the false calls and a number of missed flaws. Unfortunately, there were still inconsistent misses on the large flaws as well.

3.2 Testing with Full Weld B-Scan and Comparison with Human Performance

Instead of the previous training and testing, the ultrasonic B-scan image was divided into 48-sample-wide windows with a step of one sample and shown to the model in consecutive order moving from left to right. The centre line of the image would be highlighted as green if prediction would exceed over 50%. The results for VRR data performance can be seen in Fig. 6 for three different training methods: (a) model trained without smallest and largest flaws, (b) model trained without the 3 mm and 17 mm flaws and (c) model trained with only the largest flaws left in the training data. The grey colored prediction shows the centre line of the model window with size 48×112 samples (window width 96 mm).

When the model was presented with similar ultrasonic data as for humans, Sample 8 from [22], the model kept its performance well and acted predictably. When the model came across an edge of the flaw and moved forward it kept detecting it until the window had completely moved past the flaw. The exception can be seen in (a) and (b) in Fig. 6, where the 13.6 mm flaw is detected only by the edges of the flaw and not at all in the middle of the flaw. This might be due because when the window is on top of the large flaw and the training set has not consisted of enough large flaws over the size of the inspection window, the model cannot detect the flaw in those areas. In Fig. 6c, where the training data set has contained only the large flaws, the model is capable of detecting the large flaws in the middle as well, granted the detection is easy as the model has been taught with the same flaw type.

When considering human performance, all human inspectors found all the large flaws, but the two 1.6 mm and 2.4 mm flaws were missed by a couple of inspectors. Those misses might have been caused by interpreting the flaw as not large enough to indicate a flaw. Model (a), which had no 2 mm and 26 mm flaw types in training, barely found the 1.6 mm flaws. Case (b) in Fig. 6 which was trained without the 3 mm and 17 mm flaw types got the perfect score on the small

flaws. Case (c), which was trained with only the large flaw types, found the 3 mm flaw type and the second 2 mm flaw type easily while having difficulties with the other 2 mm flaw types. One of the reasons for these detections is that the same flaw is presented to the model multiple times as the window moves over it. Thus, there are more opportunities for finding the right features for detection compared to the test in Sect. 3.1, where detection was based only on a single attempt. This explains why the 1.6 mm flaw gets detected partially for models (a) and (c). This shows that the model has the potential for human-level performance, as these smaller flaws had not been shown to the model before.

4 Discussion

Detection accuracy seems to be highly related to the smallest flaw size used in training. While the model is capable of finding larger flaws than it is used to train with, the detection probability decreases once the tested flaws start to be smaller than those used in the training data set. This is good for qualification purposes, as it can be shown that the model generalizes better in finding the larger flaws consistently, as they are also the most critical ones to be found. In addition, the model's accuracy can be adjusted by using the flaw size range required for the task.

While the flaws available for training were limited, certain observations regarding the flaw type could be made. When the model was trained with an EDM notch within the training set or just with the EDM notch available, the generalization of the model was better than with the solidification cracks only. The most drastic effect could be observed when training with only the 6 mm solidification crack and 6 mm EDM notch in Fig. 4g and h, respectively. The comparison of these two flaw responses in Fig. 7 shows that the flaws look completely different. Also, the model trained with the EDM notch could find the solidification cracks. Whereas the model taught with only the 6 mm solidification crack struggled to find similar cracks and kept constantly missing the same-sized EDM notch. This indicates that the model might have learned features related to the solidification crack, not the pure crack indication. While the data set size was small for the single flaw types, the two types performed completely differently with the same data size. This could also be observed when the model was trained with only the larger flaws in Fig. 4d, where the detection of the 6 mm EDM notch is clearly lower and separated from the 6 mm solidification crack, which was detected with high reliability.

When comparing the performance to the whole weld image from the VRR data, the model showed consistent performance with similar results to the initial testing. The major observation was the performance drop for the larger 17 mm flaw, especially when the model was taught with the largest

Fig. 6 VRR test B-scan image, flaw locations are 220, 350, 590, 720, 820 and 880 mm along the scan axis and at a depth of 1100 samples on the sound path. Flaw sizes are 20.8 mm, 4.8 mm, 13.6 mm, 1.6 mm, 2.4 mm, 1.6 mm virtually augmented from 26, 6, 17, 2, 3 and 2 mm flaws respectively. Flaw predictions are highlighted as grey area. **a** Model trained without the smallest 2 mm and largest 26 mm flaw. Detection of the 17 mm flaw is unreliable on the middle of the flaw but more certain on the edges, 2 mm flaw at 720 mm is barely detected. **b** Model trained without the 3 mm and 17 mm flaws. Detection of 17 mm solidification crack is purely based on detecting from the edges. **c** Model trained with only the largest 17 and 26 mm flaws. Detection of the 17 mm solidification crack is reliable, but one of the two 2 mm flaws at 720 mm seems to be difficult to find

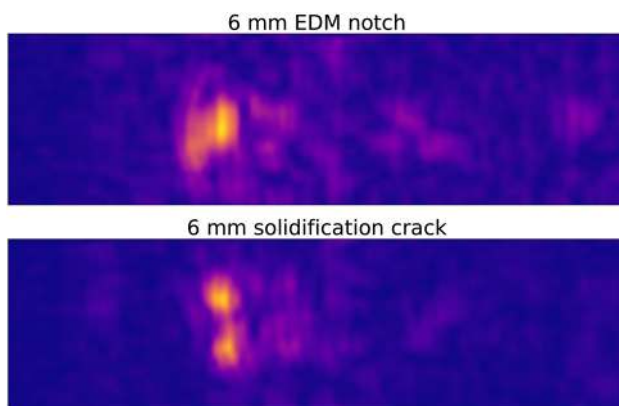
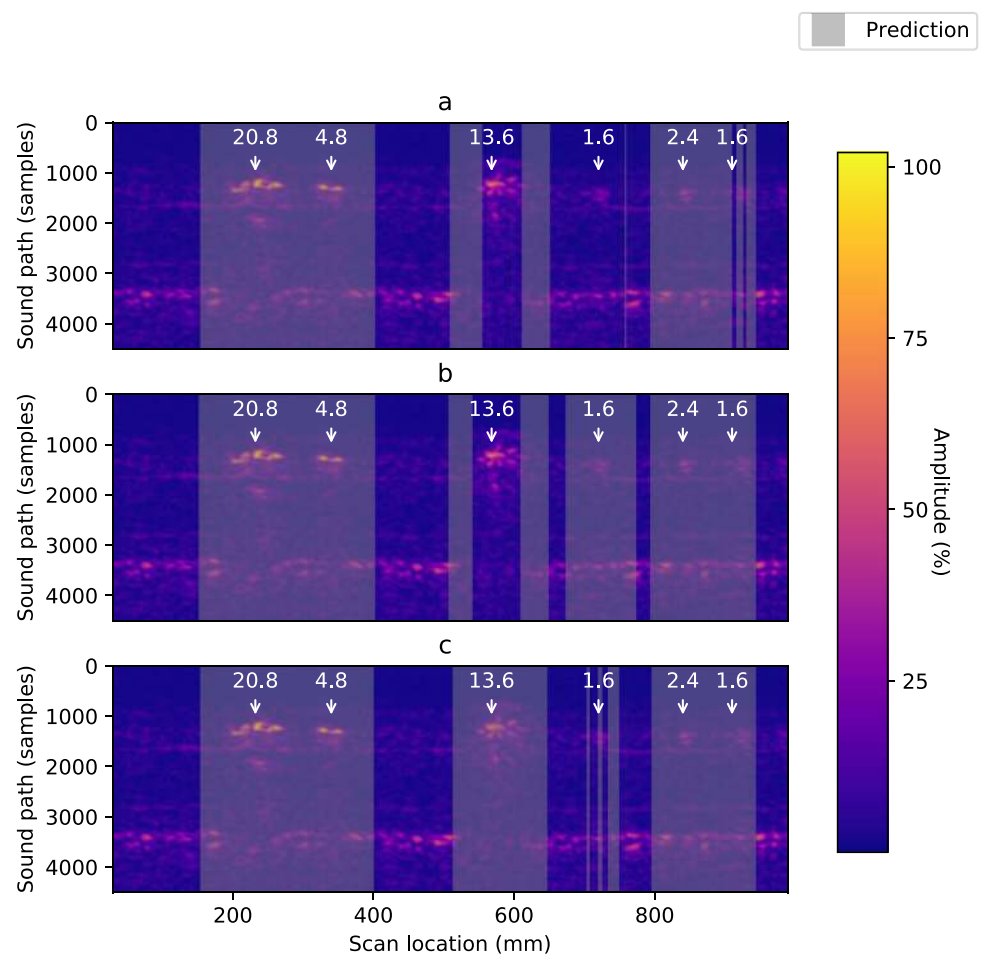


Fig. 7 Comparison between 6 mm EDM notch (above) and 6 mm solidification crack (below). The solidification crack clearly has two peaks whereas the EDM notch has a clear single peak

flaws completely left out of the training set. While the large flaw was found, the detection relied mostly on detecting the flaw from its edges. This indicates that when the flaw is large enough (i.e., wider than the observation window), the model's performance decreases drastically, as it has not experienced a similar situation when training with only small flaws. Therefore, it is highly beneficial to have larger flaws in training

to compensate for this performance decrease. This observation is consistent with the results in Sect. 3.1 where some of the 17 mm flaws were unexpectedly missed when the model was taught without the large flaws. In addition, the 17 mm solidification crack was slightly tilted, thus it gave a slightly different flaw indication than the 26 mm solidification crack in Table 1.

The effect of the smallest flaw in the data set could be seen with all real flaws. When small 2 and 3 mm flaws were included in the data set, the $a_{90/95}$ value was 1.05 mm. This result might be overly optimistic, as the flaw type was the same for testing as for training, while the flaws themselves were new through the virtual flaws. With medium 6 mm flaws as the smallest flaws, the $a_{90/95}$ rose to 3.45 mm and to 2.55 mm with the EDM notch. The better performance for the lone EDM notch can be explained for better generalization and focus on the real indication discussed above. When the smallest original flaw was 17 mm, the $a_{90/95}$ rose to 7.65 mm. The decrease in performance can be explained by missing the EDM-type flaws in large number. However, there is a clear link to the $a_{90/95}$ value and the smallest available flaw for training.

The number of false calls seemed to increase as the smallest flaw size in the data set decreased. For these data sets and flaw types in Fig. 4a, b, the threshold for the increased number of false calls seemed to be when only the small 2 and 3 mm flaws were used, resulting in 3 false calls and a rise in prediction values where flaws did not exist. However, these flaws might have been still too large and clearly visible for a proper threshold determination. The effect of the small flaw size is seen more clearly with the simulated flaws, as the false calls are decreased by 40% when the simulated flaw sizes 1 and 2 mm are excluded from the training set. These said flaws were deemed undetectable by the human eye as well.

Even though the simulated data did not provide reliable results compared to the real flaws, it needs to be noted that the simulation of the DMW case was largely simplified. It may be plausible to enhance the performance by simulating the subject in more detail, thus decreasing the false calls and improving generalization if the simulated flaw response represents the subject in greater detail.

5 Conclusions

Modern deep learning models have proven highly efficient and reliable in image recognition tasks. It is clear that the same approach can be used for NDT applications such as ultrasonic inspection. However, as these models extract the features related to detection on their own, great care needs to be taken when designing a data set for training a machine learning model for ultrasonic inspection:

- Smallest flaw size detected is related to the smallest flaw size available in the training data set.
- Flaw types may generalize differently, e.g. solidification cracks generalized worse to EDM notches than vice versa.
- Using small flaws that are nearly undetectable in training may lead to deteriorated model performance.

Acknowledgements This paper was funded by the Finnish Radiation Safety Program SAFIR2022 and VTT Technical Research Centre of Finland Ltd. Trueflaw contributed the eFlaw augmentation, and the original weld sample was provided by SQC through the PIONIC program. All contributions are gratefully acknowledged.

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

Data Availability The training data set is made available for download at https://github.com/koomas/NDT_ML_Flaw.

Compliance with Ethical Standards

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. 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>
2. Annis, C.: MILI-HDBK-1823a, Nondestructive Evaluation System Reliability Assessment. Technical report. Department of Defence (2009). [http://www.statisticalengineering.com/mh1823/MIL-HDBK-1823A\(2009\).pdf](http://www.statisticalengineering.com/mh1823/MIL-HDBK-1823A(2009).pdf)
3. Chollet, F.: *Deep Learning with Python*, 1st edn. Manning Publications Co., Greenwich (2017)
4. *Extend: CIVA NDE 2017 User Manual* (2017)
5. Garbin, C., Zhum, X., Marques, O.: Dropout vs. batch normalization: an empirical study of their impact to deep learning. *Multimed. Tools Appl.* (2020). <https://doi.org/10.1007/s11042-019-08453-9>
6. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov R.: Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*. abs/1207.0580 (2012). <http://arxiv.org/abs/1207.0580> arXiv:1207.0580
7. *Inspecta Sertifiointi Oy Finnish Methodology for Qualification of PSI/ISI NDT-Inspection Systems According to STUK YVL E.5 Scheme*, 3rd issue (2019)
8. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. *CoRR* abs/1502.03167 (2015). <http://arxiv.org/abs/1502.03167> arXiv:1502.03167
9. Kemppainen, M., Virkkunen, I.: Crack characteristics and their importance to NDE. *J. Nondestruct. Eval.* **30**(3), 143–157 (2011). <https://doi.org/10.1007/s10921-011-0102-z>
10. Koskinen, A., Leskelä, E.: Differences in defect indications of three artificially produced defects in ultrasonic inspection. In: *Baltica IX. International Conference on Life Management and Maintenance for Power Plants*, No. 106, 11 July 2013 to 13 July 2013, pp 581–602. VTT Technical Research Centre of Finland, Finland, VTT Technology (2013)
11. Koskinen, T., Virkkunen, I.: Hit/miss pod with model assisted and emulated flaws. In: *12th European Conference on Non-destructive Testing (ECNDT 2018)*, *NDT.net*, No. 8. e-J. *Nondestruct. Test.* (2018)

12. Koskinen, T., Virkkunen, I., Papula, S., Sarikka, T., Haapalainen, J.: Producing a pod curve with emulated signal response data. *Insight* **60**(1), 42–48 (2018). <https://doi.org/10.1784/insi.2018.60.1.42>
13. Li, X., Chen, S., Hu, X., Yang, J.: Understanding the disharmony between dropout and batch normalization by variance shift. *CoRR* abs/1801.05134 (2018). <http://arxiv.org/abs/1801.05134> [arXiv:1801.05134](https://arxiv.org/abs/1801.05134)
14. Masters, D., Luschi, C.: Revisiting small batch training for deep neural networks. *CoRR* abs/1804.07612 (2018). <http://arxiv.org/abs/1804.07612> [arXiv:1804.07612](https://arxiv.org/abs/1804.07612)
15. 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>
16. 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 (2018). <https://doi.org/10.1007/s12206-018-0610-1>
17. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you? Explaining the predictions of any classifier. *CoRR* abs/1602.04938 (2016). <http://arxiv.org/abs/1602.04938> [arXiv:1602.04938](https://arxiv.org/abs/1602.04938)
18. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-CAM: why did you say that? Visual explanations from deep networks via gradient-based localization. *CoRR* abs/1610.02391 (2016). <http://arxiv.org/abs/1610.02391> [arXiv:1610.02391](https://arxiv.org/abs/1610.02391)
19. Svahn, P.H., Virkkunen, I., Zettervall, T., Snögren, D.: The use of virtual flaws to increase flexibility of qualification. In: 12th European Conference on Non-destructive Testing (ECNDT 2018), NDT.net, No. 8. e-J. Nondestruct. Test. (2018)
20. Szvai, S., Bzi, Z., Dudra, J., Mhsz, I.: Modelling of phased array ultrasonic inspection of a steam generator dissimilar metal weld. *Procedia Struct. Integr.* **2**, 1015–1022 (2016). In: 21st European Conference on Fracture, ECF21, 20–24 June 2016, Catania, Italy. <https://doi.org/10.1016/j.prostr.2016.06.130>
21. 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. e-J. Nondestruct. Test. (2014)
22. Virkkunen, I., Koskinen, T., Jessen-Juhler, O.: Virtual round robin—a new opportunity to study NDT reliability (Submitted for review, 2020)
23. Virkkunen, I., Koskinen, T., Jessen-Juhler, O., Rinta-Aho, J.: Augmented ultrasonic data for machine learning. *J. Nondestruct. Eval.* (2021). <https://doi.org/10.1007/s10921-020-00739-5>
24. Wu, H., Gu, X.: Towards dropout training for convolutional neural networks. *Neural Netw.* **71**, 1–10 (2015). <https://doi.org/10.1016/j.neunet.2015.07.007>
25. Ye, J., Ito, S., Toyama, N.: Computerized ultrasonic imaging inspection: from shallow to deep learning. *Sensors* **18** (2018). <https://pubmed.ncbi.nlm.nih.gov/30405086>, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6263978/>
26. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. *CoRR* abs/1611.03530 (2016). <http://arxiv.org/abs/1611.03530> [arXiv:1611.03530](https://arxiv.org/abs/1611.03530)

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