Ultrasonics 127 (2023) 106854

Contents lists available at ScienceDirect

Ultrasonics

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

Machine learning for ultrasonic nondestructive examination of welding defects: A systematic review☆

Hongbin Sun [a,*], Pradeep Ramuhalli [a], Richard E. Jacob [b]

[a] Oak Ridge National Laboratory, 1 Bethel Valley Rd, Oak Ridge, TN 37830, USA
[b] Pacific Northwest National Laboratory, Richland, WA 99352, USA

ARTICLE INFO

MSC:
00-01
99-00

Keywords:
Machine learning
Nondestructive examination
Ultrasound
Welding defect
Model performance

ABSTRACT

Recent years have seen a substantial increase in the application of machine learning (ML) for automated analysis of nondestructive examination (NDE) data. One of the applications of interest is the use of ML for the analysis of data from in-service inspection of welds in nuclear power and other industries. These types of inspections are performed in accordance with criteria described in the ASME Boiler and Pressure Vessel Code and require the use of reliable NDE techniques. The rapid growth in ML methods and the diversity of possible approaches indicate a need to assess the current capabilities of ML and automated data analysis for NDE and identify any gaps or shortcomings in current ML technologies as applied to the automated analysis of NDE data. In particular, there is a need to determine the impact of ML on the NDE reliability. This paper discusses the findings from a literature survey on the current state of ML for the automated analysis of data from ultrasonic NDE of weld flaws. It discusses an overview of ultrasonic NDE as used for weld inspections in nuclear power and other industries. Data sets and ML models used in the literature are summarized, along with a generally applicable workflow for ML. Findings on the capabilities, limitations and potential gaps in feature selection, data selection, and ML model optimization are discussed. The paper identified several needs for quantifying and validating the performance of ML methods for ultrasonic NDE, including the need for common data sets.

1. Introduction

Periodic in-service inspection (ISI) using nondestructive examination (NDE) methods is a part of the defense in depth strategy used in the nuclear industry for providing a reasonable assurance of safety. In the United States, in-service inspection is performed under the American Society of Mechanical Engineers (ASME) Boiler and Pressure Vessel Code (BPVC), using acceptance criteria also defined in the code. The US commercial nuclear fleet is increasingly moving toward a risk-informed approach to prioritizing inspections and deferring or eliminating unnecessary inspections [1,2]. Given this trend, there is a need to ensure that NDE is highly reliable in detecting flaws in safety–critical (Class-1) components.

As computer power increases and the number of qualified NDE inspectors declines, there is an increased interest in the industry to automate the analysis of NDE data. Advances in computational power, the ready availability of cloud-based computing, and machine learning (ML) algorithms make such automated data analysis possible. As in other science and engineering applications, ML (especially deep learning [DL]) is seen as a potential solution to enhancing the reliability of NDE data analysis. Although there has been a significant increase in the number of recent publications on machine learning for NDE, there appear to be many differences in the way machine learning is applied, and there is also great diversity in the methods themselves. The industry has progressed in the use of ML algorithms to train computer systems to detect flaws from NDE data sets. For instance, steam generator tube inspections currently use automated data analysis to help human inspectors [3], leveraging the abundance of data from steam generator inspections and the relatively straightforward nature of the analysis compared to visual and ultrasonic inspections.

The rapid growth in these methods and the diversity of possible approaches indicate a need to assess the current capabilities of ML and automated data analysis for NDE, as well as to identify any gaps or

* Corresponding author.
E-mail addresses: sunh1@ornl.gov (H. Sun), ramuhallip@ornl.gov (P. Ramuhalli), richard.jacob@pnnl.gov (R.E. Jacob).

https://doi.org/10.1016/j.ultras.2022.106854
Received 22 March 2022; Received in revised form 29 August 2022; Accepted 20 September 2022
Available online 26 September 2022

shortcomings in current ML technologies. Results of future research to address identified gaps are likely to be valuable in developing the technical basis for the applicability of ML in automated data analysis for its use for NDE of nuclear power plant (NPP) components.

This article describes the findings from a literature survey on the current state of ML for ultrasonic NDE of welding defects. For this effort, it is assumed that ML includes an element of learning from data (most automated analysis methods use classical statistical techniques). Although ML is also widely applied to structural health monitoring (SHM) [4–6] and conditional assessment (CS) [7,8], this work will focus the literature review on the ML applications for ultrasonic NDE of welds. The authors also assume that the reader is familiar with ultrasonic NDE and has a reasonable familiarity with the basics of ML. Background information on ultrasonic NDE may be obtained from one of several sources (for instance, texts such as [9,10]). Background information on ML can be obtained from a number of sources (for instance, [11–13]), and a brief overview of common ML models for NDE is included. Furthermore, the broad set of ML applications in NDE that have been described in the literature required the establishment of clear boundaries for this effort. For this study, the focus was on the classification of ultrasonic measurements from weld inspections into two or more flaw categories, with the basic classification problem categorizing measurements into flaws and non-flaws. Other applications, such as image reconstruction or inspection planning and optimization, were not considered in this literature evaluation. The NDE methods of interest to this assessment were limited to ultrasonic testing (UT) with longitudinal and shear waves – including phased array probes and conventional single- and dual-element probes – applied for inspecting welds and the adjacent heat-affected zone (HAZ) for the presence of flaws. Although the ultrasonic guided wave and acoustic emission techniques were also reported in a few studies [14–17] for welding defect characterization, these two techniques are not widely utilized for in-service weld inspection in practice. Because the ultrasonic guided wave is used for long-range inspection with limited energy focused on the weld and acoustic emission is a passive sensing technique, this is not suitable for in-service inspection.

The article begins with a brief overview of ultrasonic NDE as used in the nuclear industry. It then provides an overview of the overall workflow typically applied for the ultrasonic classification problem, discusses typical ML algorithms that have been applied to this problem, and finally describes research gaps. This work is motivated by the need for ML for ultrasonic NDE in nuclear engineering, and the ultrasonic NDE technique discussed is exactly the same as that used in other industries for weld inspections. The literature review is not limited to nuclear engineering, and the findings can be applied to any applications of using ML with ultrasonic NDE for in-service inspections.

2. Ultrasonic NDE in the nuclear industry

Ultrasonic inspections in the nuclear industry typically focus on piping and pressure vessel weld joints. To analyze data, inspectors rely on a variety of factors, including knowledge of the weld and piping geometry, multiple inspection angles, and the ability to view the data from multiple orientations, such as B-scans, C-scans, and D-scans [18]. Most importantly, inspectors rely on experience and the ability to make on-the-fly decisions. For instance, weld crowns, structural supports, or insulation may limit probe coverage [19] in piping inspections requiring inspectors to distinguish between echoes from a partially insonified flaw, an embedded flaw, or a geometrical reflection. As another example, inspectors need to be able to discern shear echo responses from longitudinal responses. Longitudinal-to-shear mode-converted signals can significantly clutter an ultrasonic scan with additional echo responses when longitudinal waves are used, such as those in austenitic weld inspections. At the same time, mode converted signals are often used by experienced analysts to provide clues about flaw location and size. A final example is a need for inspectors to discern relevant signals from noise. This is particularly important in coarse-grained materials and austenitic welds, where the grain boundaries can produce coherent reflections. Inspectors will often use echo-dynamic clues to detect flaws in noisy conditions since multifaceted cracks, as opposed to geometrical echoes, typically exhibit a varying signal response that can meander or "walk" across the timebase.

NDE inspectors are initially subjected to a performance demonstration requirement, during which they must prove their ability to detect and characterize flaws in ultrasonic scans [20]. Inspectors must meet minimum requirements of flaw detection rates, length sizing, and depth sizing to maintain the ability to conduct code-specified examinations. Such performance demonstration activities are critical for ensuring inspector competence and maintaining expertise [21].

To automate the flaw detection capabilities of an experienced inspector, all of the above detection and performance demonstration issues must be considered. Because of the heavy reliance on experience and decision-making, algorithms that only use basic parameters such as signal intensity, echo dimensions, and echo locations are inadequate. Even so, such algorithms can be used to reduce the burden on the inspector by screening out spurious signals so the inspector can then focus on potential flaw signals [22]. However, the diversity in analysis methods and algorithms generally requires insights into ultrasonic testing, and the knowledge of algorithms themselves and the results are highly dependent on the operator's experience.

3. Machine learning for NDE: Review methodology

The literature search used scientific indexing services such as Web of Science, Science Direct, Scopus, Google Scholar, and the Department of Energy (DOE) Office of Scientific and Technical Information (OSTI). Keywords included ultrasonic nondestructive testing/evaluation (NDT/NDE), weld defect/flaw, NDE automated analysis, machine learning, and deep learning for NDE. Out of about 200 articles identified through this search, about 135 were determined to be distinct and relevant to this article. Articles written by the same authors and substantially similar in approach and results were represented using a single representative article in this review. The search, though extensive, was not comprehensive and may have missed some articles of relevance. However, the information extracted from this analysis was still expected to be broadly applicable given the specific questions that were being considered. Specifically, the literature review assessed

- data sources, data completeness, and data management,
- methods and associated parameters (including preprocessing and ML),
- sensitivity of results to parameters, sources, and level of uncertainty,
- metrics to assess ML performance,
- methods for verification and validation, and
- bias in data and results.

Given the rapid advances in ML, it is likely that the number of publications describing the application of these techniques to ultrasonic NDE will continue to grow in the near future. The review documented in this article was a point-in-time assessment that attempted to identify research needs that are expected to continue to be relevant to the ML for the ultrasonic NDE community.

The literature assessment did not check for the appropriateness of the ultrasonic NDE procedures used to generate the data (for instance, the appropriateness of the selected frequency or probe type for the problem). Given the nuclear industry's rigorous approach to personnel and procedure qualification, it was assumed that appropriate procedures will be used for the inspections as part of any future ML applications to ultrasonic NDE data analysis in the nuclear industry.

A summary of the inspection setups from the surveyed articles is shown in Fig. 1. Fig. 1a shows the approximate percentage of articles

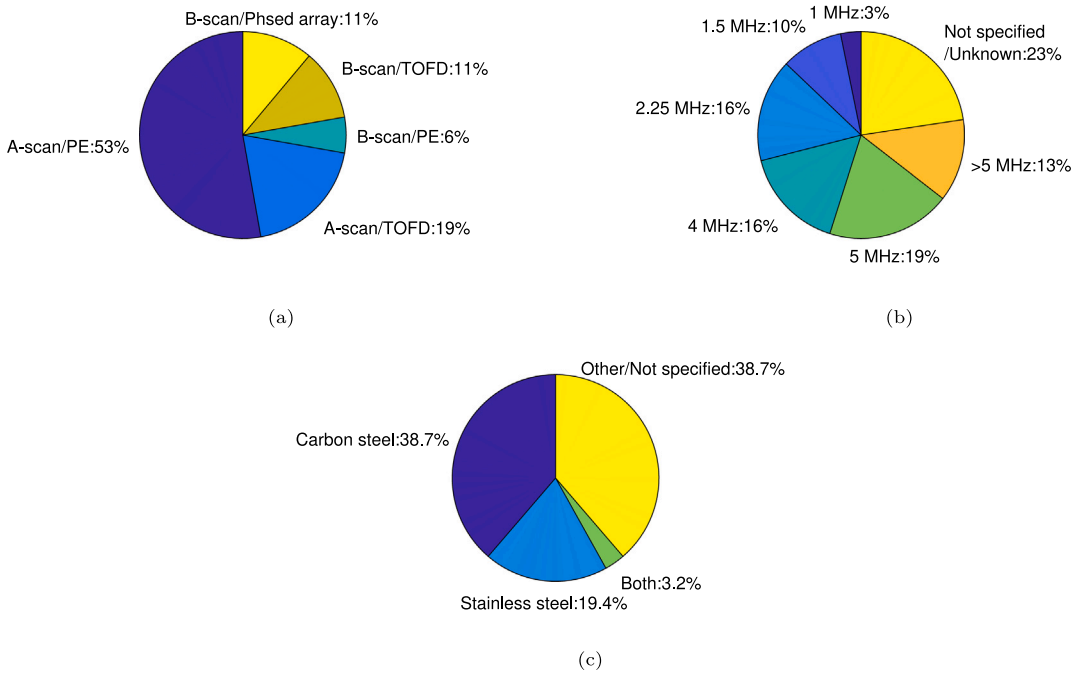

Fig. 1. Summary of percentage of articles reviewed, sorted by: (a) The type of ultrasonic data, (b) the frequency used for inspection, and (c) the specimen material.

that used various combinations of signal (A-scan vs. B-scan) and inspection approach (phased array vs. pulse-echo (PE) vs. Time-of-Flight Diffraction (TOFD)). Fig. 1b summarizes the inspection frequencies used in these studies, and Fig. 1c shows the materials that were inspected. These data indicate a relatively broad application of ML methods, covering most scenarios likely to be encountered in a typical nuclear plant inspection.

4. Machine learning for ultrasonic NDE

Rapid advances in computational techniques and data science have led to the development of "intelligent" analysis techniques that rely on ML. Harley and Sparkman [23] summarized the ML models used for solving NDE problems at a high level and proposed three significant challenges, whereas Wunderlich et al. [24] summarized the applications of ML in NDE. Bowler et al. [25] reviewed ultrasonic sensing and machine learning for various industrial applications, including weld inspection. It is feasible that "intelligent" algorithms that rely entirely on ML may eventually replace the human inspector if enough empirical teaching data sets are available [26]. However, there is likely no one-size-fits-all approach to "intelligent" automated data analysis given the diversity in flaw types, inspection techniques, inspection parameters, and analysis algorithms and processes.

Currently, there are no "intelligent" UT weld inspection data analysis software packages commercially available for piping inspections. Several software packages are available that have automated analysis capabilities, but their algorithms are essentially based on signal intensity and rely on user-defined parameters with significant limitations. For example, two commercial packages, ULTIS (TD NDETM, Montreal, Quebec, Canada) and NDTkit (Testia, Toulouse, France), provide automated data analysis tools focused on the aerospace industry. ADT PRO® (VeriPhase®, Inc., Birmingham, Alabama, USA) is a data analysis package currently available for piping weld inspections and is being evaluated for application to the nuclear industry. Although it has several limitations – it currently allows only one input data format and has been vetted only for shear waves in carbon steel with a limited range of pipe thicknesses, inspection angles, and probe frequencies – such software represents important advancements in automated piping inspection analysis. Other software packages produced, for example, by WesDyne, Eddyfi, or Zetec, are available that can help filter signal responses for further analysis by an inspector. Again, none of these packages use "intelligent" approaches as yet. Several companies are developing products or services for automated analysis of NDE data using artificial intelligence and machine learning. Ondia (Ondia, Québec, Canada) is a web-based platform that can analyze the 2D and 3D ultrasonic phased array data automatically using AI. Trueflaw (Trueflaw Ltd. Espoo, Finland) developed an embedded system for automated flaw detection based on machine learning. The system is pre-trained with a combination of client data and Trueflaw's virtual cracks to obtain comparable performance with human analysts.

4.1. Ultrasonic data sets for machine learning

The data used in developing ML algorithms are critical because the performance of these data-driven methods for flaw classification depends on the information available in the data. As discussed earlier, the literature on ML for ultrasonic NDE included a variety of inspection setups, with a range of frequencies, data types, and specimen materials evaluated. As a result, the literature on ML for ultrasonic NDE had significant diversity with respect to the available data for developing and assessing the performance of ML algorithms.

Tables 1 and 2 summarize the types of flaws in the first column along with the number of measurements (signals) for each class and the total number in the second column. Common flaw types in the studies include cracks (CK), porosity (PO), lack of penetration (LP), lack of fusion (LF), and slag inclusion (SI), with data from clean specimens used as examples of no flaws (NF), which are illustrated in Fig. 2. Cracks are the most serious defect type for welded joints. When the localized tensile stress exceeds the weld or base material strength, the welded joint will crack. Porosity is a gas-filled flaw created by trapped gas. A lack of fusion flaw occurs when the base metal and the weld material do not fuse properly, resulting in a gap between the two areas. A lack of penetration flaw exists when the weld material does not completely go through the thickness. Slag inclusions form when impurities are trapped inside a weld [27,28]. These flaws could greatly lower the strength of the weld and act as stress concentrators, decreasing the structural integrity of the weld. Other welding defects

Table 1
A summary of the flaw data in literature.

Flaw classes	Instances for classes (Total)	Crack, Non-crack flaw, Non-flaw	Ref.
NF, LF, LP, PO	60, 60, 60, 60 (240)	0-180-60	[33]
CK, PO, slag line	15, 15, 20 (50)	15-35-0	[34]
LF, NF, CK	100, 100, 73 (273)	73-100-100	[35]
CK, stomata, LP, SI	60, 60, 60, 60 (240)	60-180-0	[36]
Hole, no hole	142, 140 (282)	0-142-140	[37]
Good weld, qualified weld, unqualified weld	33, 33, 34 (100)	0-67-33	[38]
Cavity, inclusion	33, 28 (61)	0-61-0	[39]
Thermal damages	40, 40, 40, 40 (160)	0-120-40	[40]
Good weld, undersize weld, stick weld, no weld	221, 30, 65, 122 (438)	0-217-221	[41]
CK, PO, SI	104, 53, 82 (239)	104-135-0	[42]
CK, PO, SI	45, 45, 45 (135)	45-90-0	[43]
CK, couterbore, rootweld	98, 73, 75 (246)	98-148-0	[44]
CK, couterbore, rootweld	107, 94, 92 (293)	107-186-0	[44]
Side drilled holes, slots, SI, PO, CK	21, 8, 13, 11, 12, 25 (90)	25-44-21	[45]
Regular belt, thick belt, large belt, PO, LF	24, 24, 24, 24, 24 (120)	0-96-24	[46]
LP, LF, PO, NF	60, 60, 60, 60 (240)	0-180-60	[47]
LP, LF, PO, NF	50, 50, 50, 50 (200)	0-150-50	[47]
External corrosion, internal corrosion, LP	180, 180, 180 (540)	0-540-0	[48]
Good weld, unsize weld, stick weld, no weld	108, 108, 108, 108 (424)	0-324-108	[49]
PO, LF, tungsten inclusion, NF	60, 60, 60, 60 (240)	0-180-0	[50]
LP, LF, NF, PO, SI	40, 40, 40, 40, 40 (200)	0-160-40	[51]
LP, LF, SI, Excess penetration	35, 32, 37, 37 (141)	0-141-0	[52]
NF, LP, SI, PO	100, 100, 100, 100 (400)	0-300-0	[53]
Failed weld, stick weld, defective weld, good weld	150, 150, 150, 150 (600)	0-450-150	[54]
LP, LF, NF, PO, SI	40, 40, 40, 40, 40 (200)	0-160-40	[55]
LF, LP, SI, PO, NF	200, 200, 200, 200, 200 (1000)	0-800-200	[56]

Table 2
A summary of the flaw data in literature (continued).

CK, LF, LP, PO, SI	121, 115, 34, 35, 35 (340)	121-219-0	[57]
CK, LF, LP, PO, SI	1200, 1150, 350, 350, 550 (3600)	1200-2190-0	[29]
Normal, pore, spatter, overlap	100, 100, 100, 100 (400)	0-300-100	[58]
Counterbore, bottom notches, bottom holes, side drill hole	1350, 675, 900, 900 (3825)	0-2475-1350	[30]
LF, LP, PO, SI, CK	16, 19, 13, 15, 12 (75)	12-63-0	[59]
CK, LF, SI, PO, LP	2899, 1196, 634, 493, 617 (5839)	2899-2940-0	[60]
CK, NF	1080, 1080 (2160)	1080-0-1080	[32]
CK, NF	5000, 5000 (10 000)	5000-0-5000	[31]
CK, NF	10 000, 10 000 (20 000)	10 000-0-10 000	[26]

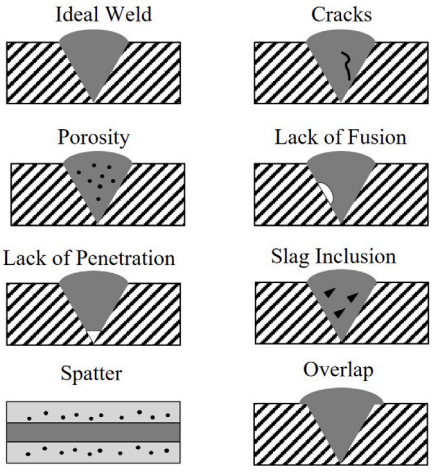

Fig. 2. Different types of welding defects.

such as spatter and overlap usually exist on the surface of the weldment and are generally considered an imperfection.

In the third column of the tables, information on the distribution of data is summarized in three categories: crack, non-crack flaws, and non-flaws. Given the difficulties with collecting and compiling inspection data, it is not surprising that most studies used 1000 or fewer signals for training and testing, with only a few studies [26,29–32] using more than 1000 signals; however, most of these studies used data augmentation methods to increase the size of the data set.

Some studies documented the number of flaws but omitted specific information on flaw types [26,29,31,57]. A few others verified the presence and characteristics of flaws in specimens using radiography to ensure that the data were correctly labeled [36,50,52,53,56]. However, most studies were silent on whether the presence of flaws was independently verified.

As discussed, many of the articles reviewed did not include detailed information on the specimens used, and several articles were based on data from a limited number of specimens (in some cases, one specimen) [26,37,53,56,59]. These studies tended to use multiple sets of measurements from the same set of flaws through repeated tests. As a result, the data used in these studies may not represent the conditions encountered in field inspection settings. In addition, ML models trained using these data with low diversity may show a low generalization capability (low classification accuracy with data not used in training), and the reported performance likely has a limited impact beyond the study. In some cases, data were generated using simulations or through "virtual flaw" techniques [26,35,61,62]. Other studies used as many measurements as the number of physical flaws present in the specimens and could increase the diversity/representativeness of the data. However, the impact on ML performance may be limited if the number of flaws is small, or the physical characteristics of the flaws are similar [34,38,43,45,52].

4.2. Machine learning workflow for ultrasonic NDE

The literature review indicated a generally applicable workflow for "intelligent" automated data analysis (Fig. 3). Broadly speaking,

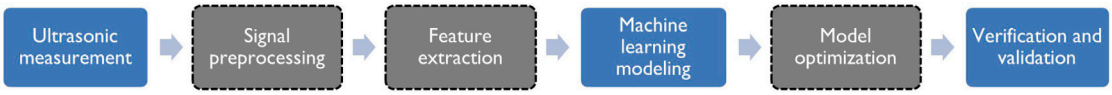

Fig. 3. Summary of workflow for applying machine learning for UT signal classification.

Table 3
Summary of the machine learning models used in ultrasonic NDE of welding defects.

Method	Features	Comments	Reference
Kernel method (e.g. SVM)	Wave parameters, statistical parameters, and wavelet features	Mostly used for two-class classification, needs feature extraction first	[33–39,63–68]
Ensemble method (e.g. random forest)	Wave parameters and statistical parameters	Higher resistance to the overfitting and higher generalization, needs feature extraction first	[40,41,69–73]
Shallow neural network	Wave parameters, statistical parameters and wavelet features	Most common model for ultrasonic NDE of welding defects, needs feature extraction first	[42–56,68,74–86]
Fully connected deep neural network	Wave parameter, statistical parameter, wavelet features, time-/frequency-domain signals	Accept time-/frequency-domain signals as feature input	[29,57,87,88]
Convolutional neural network	Wave parameter, statistical parameter, wavelet features, time-/frequency-domain signals, B-scan image	Mostly used for image classification, accept time-/frequency-domain signals as feature input	[26,29–32,60–62,67,86,89–97]
Unsupervised learning	Clustering [44,98–100], PCA [35,53,99,101,102], KNN [68,99,102], Anomaly detection [101], Unsupervised neural network [52,103]		
Multiple learning models	SVM + neural network (NN) [35], Decision tree + random forest [41], Fuzzy classifier + NN [78], NN + Self-organizing map + Probabilistic NN [52], Shallow NN + DNN [57], MLP + k-Nearest Neighbor + NN + decision tree [82], DNN + CNN [29], CNN + SVM [32], PCA + KNN [102], PCA + KNN + K-mean [99], K-mean clustering + autoencoder [104], K-clustering + Gaussian mixture modeling + Mean shift clustering [100], BPN + RNN + LSTM [59], CNN + RNN [92], SVM + MLP + KNN [68]		

ultrasonic NDE data are usually preprocessed to eliminate any outliers, address missing data issues, and reduce noise. Features—essential signatures that improve discrimination between flaw and non-flaw data—are often extracted and used as the input to the ML algorithm. A portion of the available data (called the training data set) is used to train the ML model. The training process optimizes the ML model parameters and hyperparameters to maximize classification accuracy or other relevant performance metrics. The training step is generally followed by a validation step, where a data set that the ML algorithm has not encountered is used to ensure that the ML model does not overfit the training data set. In ML, overfitting refers to the condition where the ML model essentially memorizes the training data and is unable to generalize its learning to other similar data sets. The training and validation stages may need to be done several times to find the best ML model parameter set. A separate test data set is then used to assess the true generalization performance of the trained ML model.

There is sufficient room within this general workflow to accommodate variations. Several of the stages in Fig. 3 are applied optionally (dashed boxes), whereas others may use the same stage multiple times to meet different objectives. For example, Refs. [33,36,48] used two stages of the feature extraction and ML algorithm blocks; the first stage was used for screening between flaws and non-flaws, and the second stage was used for identifying the type of flaws. Modern deep learning algorithms often skip the feature extraction stage, preferring to let the algorithm extract an implicit representation of relevant features during the learning phase.

The step of feature extraction usually requires a prerequisite of signal preprocessing. For A-scan signals (e.g., pulse-echo and TOFD), these signal parameters can be extracted from the time domain signal: arrival time, wave velocity, wave attenuation, maximum amplitude, rise time, and fall time. Features related to the echoes may be extracted if multiple echoes exist. In the frequency spectrum, peak frequency, maximum amplitude, mean frequency, and the number of maxima can be used. Statistic parameters could be extracted from both time-domain signal and frequency spectrum. The preprocessing for A-scan signals includes amplitude normalization, time scaling, noise elimination, spectrum processing (e.g., fast Fourier transform [FFT], and discrete wavelet transform [DWT]), etc. With DWT, the wavelet feature is another widely used feature from ultrasonic signals. For B-scan images, textural features are the most straightforward features, including contrast, correlation, energy, entropy, co-occurrence matrix, etc. Similar to the A-scan signals, statistic parameters can be directly applied to B-scan images with pixel values. Wavelet features can also be used for B-scan images with 2D wavelet transform. Physical parameters such as welding defect size and position information can be extracted from the B-scan image. The B-scan image can also be used as the input feature directly for deep learning models such as convolution neural networks. The B-scan image also needs preprocessing, such as normalization, scaling, de-noising, converting to gray-scale, and segmentation before feature extraction. The details of the feature extraction and preprocessing are explicitly discussed in Section 5.1

Table 3 summarizes the major ML models investigated in the reviewed papers, along with the features used. These ML models are briefly introduced and reviewed here: kernel, ensemble, neural network, and unsupervised learning methods. The support-vector machine (SVM) was a representative method of the kernel machine model for pattern analysis, and it was widely used for binary class classification. For multiple class classification, it might require particular strategies such as one-against-one [64], one-against-all [34], and binary decision tree strategies [82]. The basic principle of the SVM method is to find a hyperplane to separate the data with a maximum margin. For a linear model with data set (x, y), the hyperplane $f(x)$ is expressed as

$$f(x) = wx + b = 0, \tag{1}$$

where w is the weight vector and b is the bias term. The hyperplane should subject to the following constraint:

$$y_i f(x_i) = y_i(wx_i + b) \geq 1, \tag{2}$$

whereas the cost function is:

$$J = \frac{1}{2}\|w\|^2. \tag{3}$$

The above constraint and cost function is for separable data. If the data are not separable, then a soft margin is used with a cost function:

$$J = \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{m}\xi_i \tag{4}$$

with constraint:

$$y_i f(x_i) = y_i(wx_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \tag{5}$$

where ξ_i is a positive slack variable that indicates that the sample training allows a small number of erroneous samples, and c is the penalty parameter. The SVM model has several hyperparameters: the kernel function, kernel scale, and penalty parameter. Linear kernels [64], Gaussian radial basis function (RBF) [35,37,105], polynomials [34,65], and sigmoid functions are among the most common kernels functions. The kernel scale defines how far the influence of a single training example reaches. The penalty parameter (PP) controls the trade-off between minimizing and maximizing the classification margin and training error. The main features used for SVM were wave parameters, statistical parameters, and wavelet features for ultrasonic NDE.

The ensemble method is a machine learning algorithm that combines the results from multiple learning models. It can provide better performance than a single model due to the diversity of the models used [106–108]. The variety of the models allows different decision boundaries and different types of errors by using different model parameters and training data sets [109]. The random forest method is a typical ensemble method of Bootstrap aggregating, which uses multiple decision trees for modeling. Nunes et al. [40] used optimum path forest and Bayesian classifiers to classify the welded steel specimens with different thermal aging damage. Camacho-Navarro et al. [110] combined decision tree algorithms with bagging to detect the damage in carbon steel pipelines. Polikar et al. [69] developed an incremental learning algorithm based on ensemble learning models for the identification of ultrasonic signals from stress corrosion cracks in the presence of counterbores and weldroots. Martín et al. [41] developed a framework using a tree-based learning model to assess the welding quality of resistance spot welding. Wave parameters and statistical parameters were used as the features for the ensemble method.

An artificial neural network is a connected graph with multiple layers, and each layer has multiple nodes (neurons). Inputs to the neural network can be time-domain features [42,47,49,54,85], frequency-domain features, wavelet features, time-domain signals, and frequency spectrum. The shallow feed-forward neural network has a relatively simple structure, with one input layer, one to two intermediate (hidden) layers, and one output layer (see Fig. 4(a)). Most studies use shallow neural networks for welding defects classification with only one hidden layer or two hidden layers. The number of neurons in the hidden layers can be optimized to achieve the best model performance. Another hyperparameter is the activation function, which defines the output of the node given the input or a set of inputs. Typical activation functions include the Gaussian function [42], sigmoid function [43,74], hyperbolic tangent [46,49,50,53], and the radial basis function (RBF) [44,52]. The learning algorithms applied include gradient descent [56], the Newton method, Quasi-Newton method, and Levenberg–Marquardt [46,49,53,54,79].

Deep learning models such as deep neural networks have been applied to NDE in the last few years for automatic defect characterization and are mainly utilized in problems with 1D or 2D features such as radio tomography image [111,112], infrared thermography images [113,114], and ultrasonic A-scan signals and B-scan images. A fully connected DNN contains more input nodes and hidden layers than a shallow neural network (see Fig. 4(b)) and can accept time- or frequency-domain signals as input features [29,57]. However, model regularization should be introduced to prevent overfitting. Regularization methods proposed for fully connected neural networks include dropout, early stopping, l_1 and l_2 regularization, and max-norm [57] regularization. Dropout is the most popular regularization method [115], with one or more dropout layers in a typical DNN (see Fig. 4(b)).

A convolutional neural network (CNN) is another type of DNN with convolutional layers and is widely used in visual imagery analysis. It requires relatively little preprocessing of the input data compared to other image classification algorithms. CNNs have been used for weld defect classification with time-domain signals and B-scan images as the inputs [29,30,32,67,89,90]. In a CNN, the convolution layer and a pooling layer act as the feature extraction layers, and fully connected layers are for classification (see Fig. 4(c)). This architecture allows CNNs to concentrate on low-level features that are then assembled into higher-level features in later layers. According to Zhang et al. [116], the filter size of the first convolution layer could be kept large to improve the performance when the signal is noisy. The stride size controls the overlap of the receptive field, and a significant stride size represents a smaller overlap in the receptive field. Munir et al. [29] used a large stride size to replace a pooling layer. The rectified linear unit (ReLu) is a typical activation function used for the convolutional layer, whereas the exponential linear unit (Elu) could be selected for a convolution layer to avoid the vanishing gradient problem [117]. A pooling layer is used to subsample the input, reduce the computational load, and avoid overfitting, but it introduces more complexity [115]. Fully connected layers in CNNs are similar to those used in a DNN for classification function, which could be replaced by other classification models such as SVM [32,118].

Unsupervised learning works with unlabeled data, extracting the feature, grouping the data, and providing the output with minimum human supervision. K-means clustering is the most common clustering method for partitioning the feature space. Polikar et al. [44] used k-means clustering to classify welding defects and found that the performance was significantly poorer than the performance of a neural network. Additionally, the k-means clustering result depends on the number of clusters used and the initial cluster centers selected. Murta et al. [99] used k-means clustering to classify simulated TOFD ultrasonic signals for welding defects, and the number of clusters was decided using the Silhouette index [119] and Davies–Bouldin index [120]. Virupakshappa et al. [100] used several clustering methods for ultrasonic flaw detection, including k-mean clustering, Gaussian mixture modeling (GMM), and mean shift clustering (MSC). Principal component analysis (PCA) is usually used for dimensional reduction and feature selection combined with other ML methods such as SVM [35], neural networks [53], k-means clustering [99], and anomaly detection [101]. PCA converts the original dataset to a low-dimensional dataset that is linearly uncorrelated, usually making the classification task easier. Anomaly detection is another unsupervised learning method that identifies anomalies from normal data. Cassels et al. [101] combined PCA and anomaly detection to detect defects. A self-organizing map (SOM) is a type of unsupervised neural network, and it produces a low-dimensional, discrete representation of the input space of the training samples. Seyedtabaii [52] used an SOM with a multilayer perceptron to classify defects. Martín et al. [103] used PCA and SOM to classify defects on rolled steel based on visual inspection data. These unsupervised learning models were usually used with other models for flaw classification. The reviewed literature showed more examples of using multiple learning models than those using a single learning model. The last two rows of Table 3 summarize unsupervised learning models and studies using multiple learning models.

The overall workflow and the diversity in ML methods seen in the literature indicate that a wide variety of methods have been explored, including different combinations of ML models and feature selection approaches. Loss functions for training the ML appear to be largely standard loss functions (mean square error and cross-entropy loss). Performance metrics reported include classification accuracy, true positive and false positive rates, cross-entropy loss, efficiency product, and occasionally, receiver operating characteristic (ROC) curves (assessing sensitivity to model parameters). Results are generally reported on a separate (previously unseen by the ML model) data set, also referred to as a test data set.

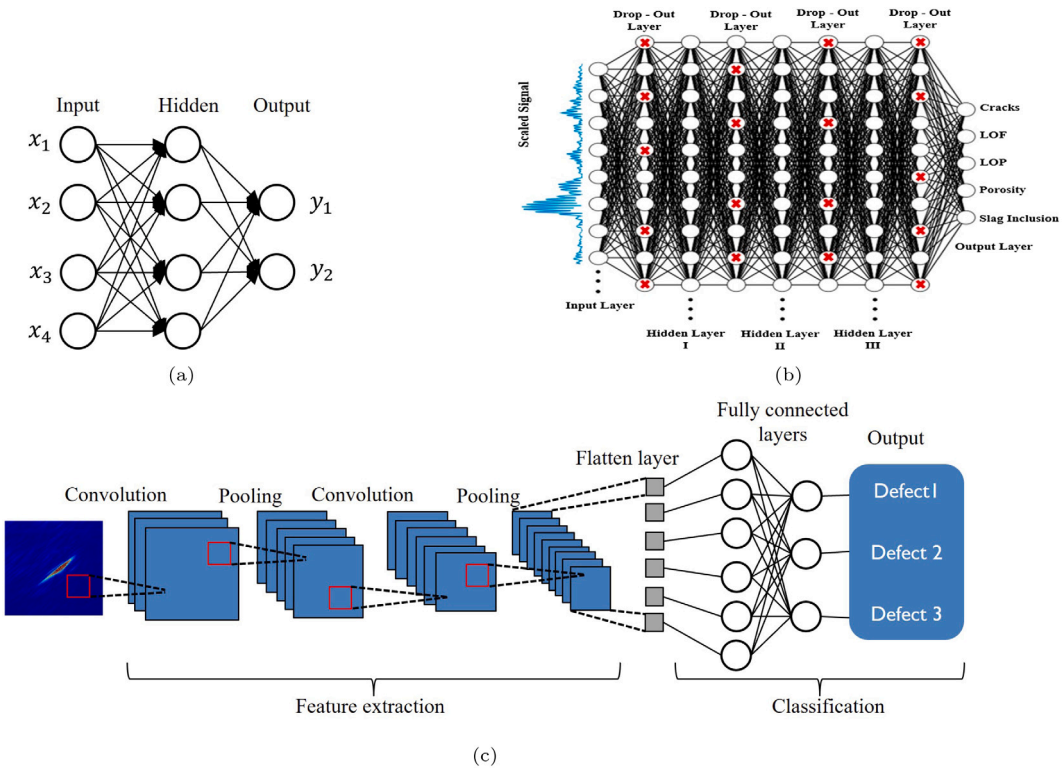

Fig. 4. (a) Example of a shallow neural network, (b) fully connected deep neural network (Reproduced from Munir et al. 2019 [29], with permission from Elsevier), and (c) convolutional neural network.

4.3. Model performance

The reported classification performance from the reviewed literature was analyzed to determine whether factors contributing to the performance could be easily determined. Fig. 5 summarizes the reported performance, represented by the classification accuracy as a function of several quantities. Fig. 5a–c show the performance as a function of (a) the ultrasonic measurement type (A-scan vs. B-scan), (b) the learning model, and (c) the feature type.

The reviews and summaries in Fig. 5 highlight several aspects of the application of ML for ultrasonic NDE classification. First, literature on ML classification of weld inspections seems to be biased toward the use of A-scans. Second, the data set sizes (as described earlier) are relatively small: most data sets are on the order of about 1000 signals or less. Data augmentation was a factor in the larger data sets used.

With these underlying data-related constraints, the classification accuracy (fraction of signals correctly classified) appears to range from about 50% to 100%. There does not appear to be any specific trend in classification accuracy with respect to the use of A-scans or B-scans (Fig. 5a). Furthermore, there are no distinguishable trends for the number of signals used in the training or the classification accuracy for A-scans or B-scans. This appears to be indicative of other factors potentially also playing a role in the reported accuracy.

An examination of Fig. 5b shows a potential relationship between the reported performance and the ML model. Specifically, techniques such as SVM and shallow NN – which tend to have fewer trainable parameters than the deep neural network or other deep learning models – seem to demonstrate better classification performance with smaller data sets. As expected, the DNN models (including fully connected NN and CNN) that require large data sets report better performance as the number of training examples increases. In addition, Fig. 5c indicates that studies with images or image-related features appear to show higher classification accuracy. However, the data used in these studies seem to be limited, and it is difficult to assess whether the results were caused by data bias. Furthermore, the data sets used in the different studies evaluated herein differ in terms of the materials inspected and ultrasonic inspection parameters, making it difficult to perform robust comparisons of the ML performance.

A slightly different perspective on the classification accuracy can be seen in Fig. 6, which shows the true positive rate or *TPR* versus the false positive rate or *FPR*. The metrics *TPR* and *FPR* are defined as:

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}, \tag{6}$$

where *TP*, *FN*, *FP*, and *TN* are true positive, false negative, false positive, and true negative, respectively. The number of articles reporting this data was limited: most articles reported only the classification accuracy for all classes combined. Ideally, classification performance should be near the upper left corner (high TPR and low FPR), and at least a few studies reported performance in this region. Similar plots of performance, when reported as a function of a decision threshold, are called receiver operating characteristic (ROC) curves [121–124]. ROC curves represent the plots of *TPR* vs. *FPR* at different decision thresholds and provide a simple approach to comparing different ML models and assessing any trade-offs associated with obtaining a higher *TPR*. When comparing ROC curves, a higher *TPR* at the same *FPR* indicates better classification performance. The area under the ROC curve denotes area-under-curve (AUC), and a higher AUC indicates a better performance. Fig. 7 shows an example of three ROC curves for a neural network model using the features extracted from three methods (discrete Fourier transform, discrete cosine transform, and discrete wavelet transform) [53]. The AUC of the ROC curve using the Fourier transform was the highest among the three, which indicates that this model performance was the best. Interestingly, very few articles appear to explicitly calculate and present the ROC curves, making it difficult to compare and contrast performance.

Taken together with the results summarized in Fig. 5, it is apparent that ML algorithms are capable of high classification accuracy, high TPR, and low FPR. However, the lower performance numbers reported in several articles also highlight the impact that other factors not explicitly reported may influence ML performance. These factors include,

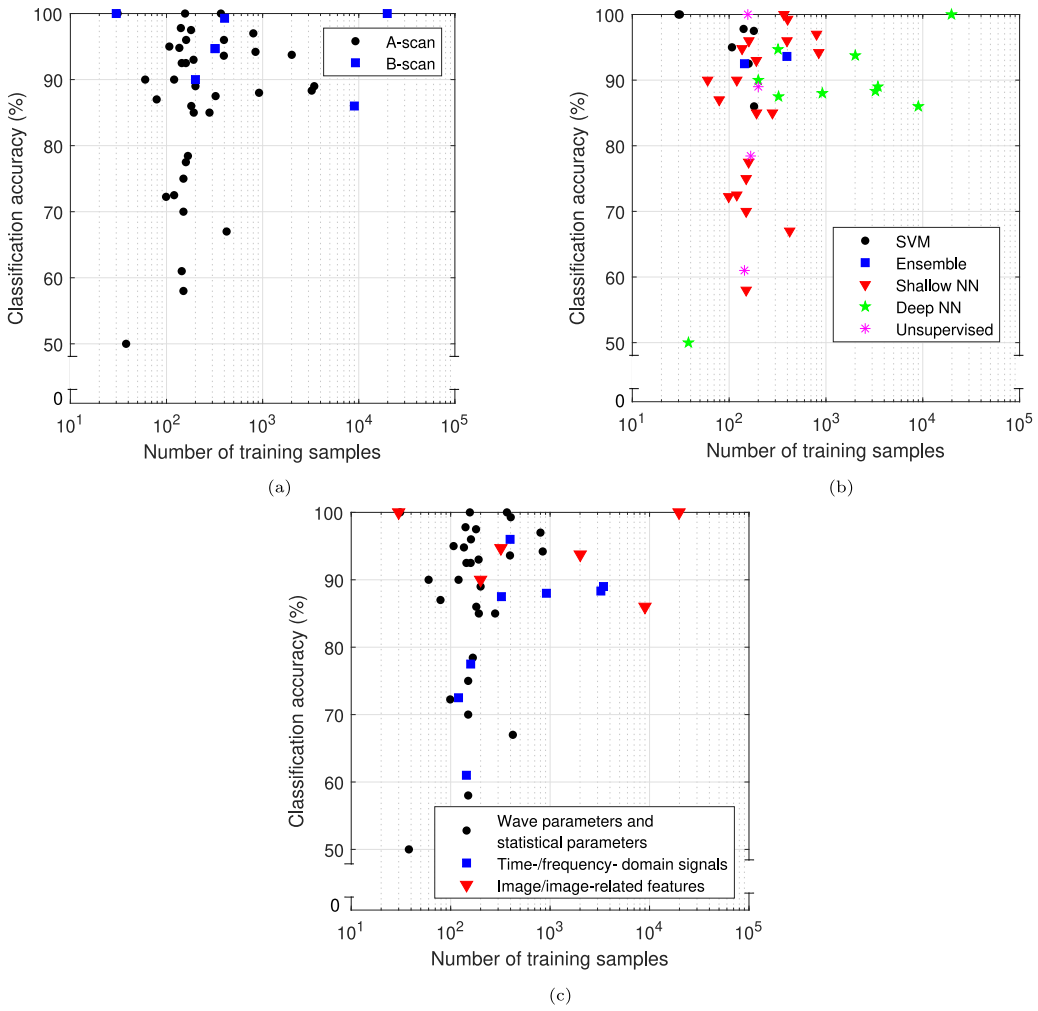

Fig. 5. Number of training samples vs. classification accuracy based on: (a) the type of ultrasonic measurement, (b) the learning model used, and (c) the feature type used.

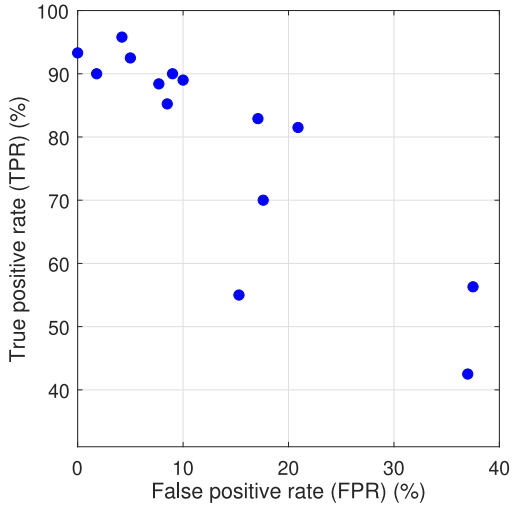

Fig. 6. A summary of classification accuracy, reported using true positive rate and false positive rate.

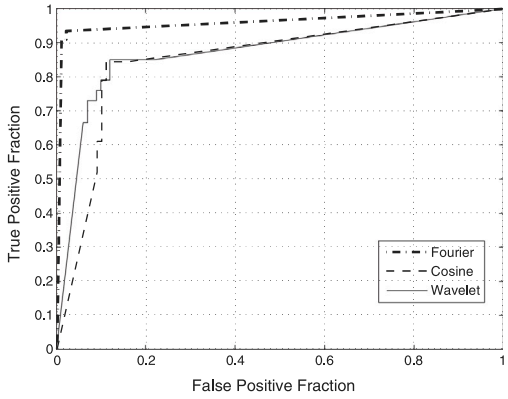

Fig. 7. ROC curves for a neural network model using features extracted from three different methods (Reproduced from Cruz et al. 2017 [53], with permission from Elsevier).

for instance, optimization of the ML algorithm parameters as well as the parameters of the learning algorithms (hyperparameters); data set characteristics including data unbalance and bias; feature extraction and selection methods used, and the level of validation performed using independent data sets to evaluate performance. The performance is also impacted by the inspection technique used and the classification objective. Often, most poor performance appears to be caused by the low quality of input features and the use of inadequate ML models. It is also noteworthy that the results do not highlight any one ML model as the best for all conditions, indicating that a certain amount of trial and error may be necessary to find the best ML model. Model parameter

and hyperparameter tuning may also be necessary to obtain the best ML model for a given set of conditions.

As an example of the impact of ML model parameter impact, Margrave et al. [45] showed that the neural networks with fewer nodes in the input layer had a poorer performance. The data reduction process to apply the signals to the NN inputs appeared to result in information loss, leading to poor performance. Similarly, combining data at two inspection frequencies without any transformations to match the characteristics of the two groups of signals resulted in poor classification accuracy (<40%) [44]. In this case, scaling the two groups of signals appropriately along the time axis significantly improved the classification accuracy (>80%). As an example of a potential mismatch between inspection technique and classification objective, Veiga et al. [47] reported that the classification accuracy of lack of penetration (LP) was around 55% when using pulse-echo signals (opposite surface) while that using TOFD signals was around 90%. Similarly, poor performance may occur when using only a part of the ultrasonic signal (diffraction wave part of TOFD, for instance, [56]) or if the data and selected ML model are mismatched somehow [29].

5. Machine learning for NDE: Takeaways

This section discusses the key findings from the literature assessment, organized into the following areas: selecting relevant features, data appropriateness, ML model and model hyperparameter selection, and verification and validation (V&V).

5.1. Feature extraction and selection

The literature indicates that ML can be successfully applied to A-scans, B-scans, D-scans, and various feature extraction methods can be used to meet the analysis objectives. One-dimensional time-series data (A-scans, e.g., Fig. 8a) were the most common form of data used in the analysis, though B-scan images (e.g., Fig. 8b) were also used—especially for phased array analysis [26,31,97,101] and TOFD (including TOFD time domain signal and TOFD D-scan image, e.g., Fig. 8c) [34,55,81,83]. However, the presence of higher noise from grain structure in TOFD images [125] may require the application of denoising or post-processing [83].

The review indicated the applicability of a number of different types of feature extraction methods based on the type of ultrasonic measurement. Various feature extraction methods that may be applicable are summarized in Table 4. Before applying feature extraction techniques, the data may have been preprocessed to normalize (scale the data to a standard range) and reduce noise (denoising). Other filters to enhance the signal-to-noise ratio (SNR) may also have been applied.

Feature extraction methods discussed in the literature vary by signal type (A-scan vs. B-scan) and include:

- ultrasonic signal parameters (e.g., wave velocity, attenuation, rise/fall time, amplitude),
- flaw size and position information extracted using image processing algorithms from B-scan or other image representations of the ultrasonic NDE data [78],
- features related to echoes [41,49] (e.g., height of one specific echo usually normalized to the first echo, time difference between two echoes),
- frequency domain parameters (e.g., peak frequency, maximum amplitude, mean frequency, the number of maxima) [42,54],
- time or frequency domain statistical parameters (e.g., mean amplitude, standard deviation, root mean square, summation of the absolute values, total energy, skewness, Kurtosis) [43,55,64,98],
- wavelet parameters [34,44,50,52], and
- textural features from B-scans or other image representations of the data (e.g., contrast, correlation, energy, entropy, co-occurrence matrix) [78].

With the wide availability of DL models, the time-domain signal or frequency spectrum can also be used directly as the input to the ML algorithm, thereby bypassing the feature extraction and selection step. This may be done for data represented as time-series (A-scans) and as images (B-scans). In the case of image representations, CNNs are popular for analysis, though the B-scan image may need additional preprocessing, such as grayscale conversion and segmentation, to further improve SNR.

It is important to note that although the use of certain features may help increase the accuracy of data analysis, too many features or those with little discriminatory information can have a harmful effect on the classification performance. The inclusion of such features can lead to overfitting and poor generalization of the ML algorithm, in which the performance on data not used in the training (learning) process is poor. Though this effect is partly an issue of inappropriate data selection for the training, it is also partly caused by poor feature selection.

Ideally, good features have a strong correspondence with the output data/label, and these features show high independence. Redundant and irrelevant features may have correlations with each other and weak relevance with the output. Keeping good features and removing the irrelevant and redundant features can make the learning model simpler and more robust. Therefore, the selection of appropriate features for input is critical for building an efficient learning model with a high model performance. Selection approaches used in the NDE literature range from manual selection [50,56] to the use of principal component analysis (PCA) [35] and various statistical tests to identify the most relevant features [53].

Feature selection is also of importance in other ML applications, and it appears that a broader set of techniques has been explored. These approaches include the use of statistical tests (such as Pearson's and Spearman's coefficients, the chi-square test, mutual information, the ANOVA (Analysis Of Variance) correlation coefficient, or Kendall's rank coefficient) to score and rank individual features [130]. Alternatively, the "wrapper method" measures the performance of different feature subsets by repeatedly training and testing the model [131] using a given feature set. Based on the performance for each feature subset, the method will add or remove specific features during each training run according to their contributions to the model, and the process repeats until specific performance goals are met [132]. Alternatively, feature selection can be embedded into the model training process [133] through the use of regularization parameters that weigh the contributions of each feature. Other methods, such as the random forest method, can provide the importance of each feature during the training process through quantities such as the mean decrease in impurity and accuracy.

5.2. Data selection

5.2.1. Sample size determination (SSD)

The literature suggests that ML performance is mainly reliant on the size and quality of the data sets used in the ML model training [134,135]. The question of sufficient sample size and quality, including the distribution of data across multiple classes were not explicitly reported in the ML for ultrasonic NDE literature. This is not merely an academic question; creating a larger than the necessary data set can be time-consuming and costly. However, the lack of information on this question is not an indication that the reported performance of these algorithms is lacking; rather, it indicates that there may be room for further improvement if the sample size and quality turn out to be less than adequate.

Even though the question of sample size and quality are unanswered in the ML for NDE literature, approaches to determine the data quantity are reported in the general ML literature. These include empirical methods such as the learning curve method and progressive sampling [136–138]. Both approaches empirically calculate the model performance as a function of sample size. Whereas progressive sampling uses an increasing series of sample sizes and stops when

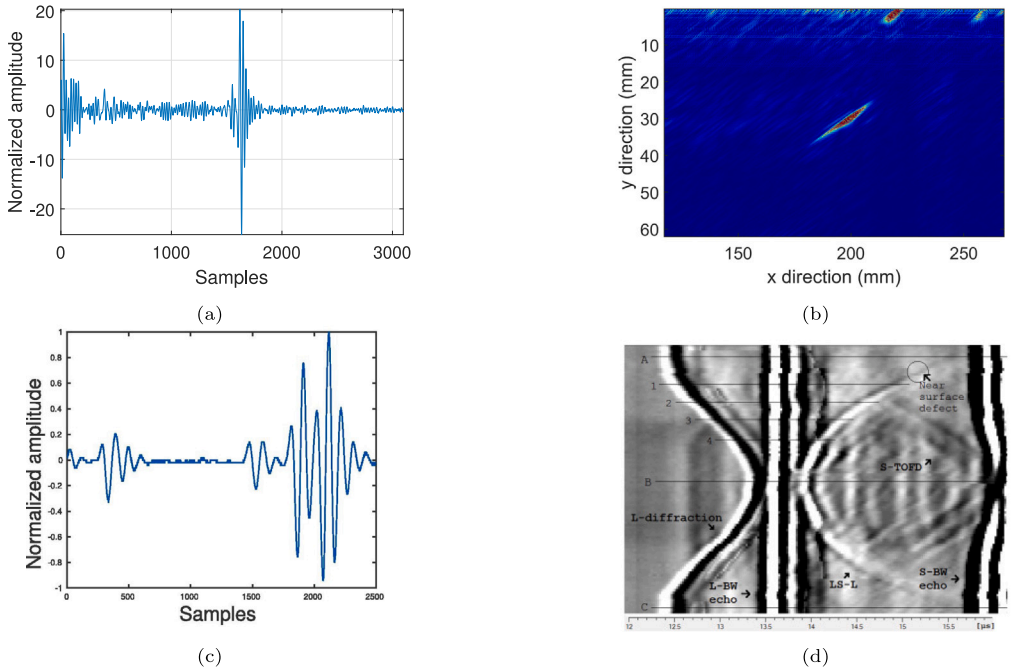

Fig. 8. (a) A-scan signal acquired by a 2.25 MHz shear transducer on a crack flaw, (b) B-scan image of a crack flaw, (c) TOFD time-domain signal collected using 5 MHz transducers on a lack of fusion flaw (Reproduced from Silva et al. 2020 [56], with permission from Elsevier), (d) TOFD D-scan image collected using 5 MHz transducers on a notch defect (Reproduced from Yeh et al. 2018 [126], with permission from Elsevier).

Table 4
Summary of feature extraction methods in the literature.

Data source	Feature type	Feature parameters	Preprocessing	Reference
A-scan signal	Wave parameter	Velocity, attenuation, rise/fall time, amplitude, peak frequency, mean frequency, number of frequency maxima, etc.	Normalization, windowing, denoising, FFT for frequency-domain parameters	[39,41–43,49,54,56,84]
	Statistical parameter	Mean amplitude/frequency, root mean square, standard deviation, summation of absolute values, total energy, Skewness, Kurtosis, etc.	Normalization, windowing, denoising, FFT for frequency-domain parameters	[32,43,51,53,54,64,70,74,79,82,98,127,128]
	Wavelet features	Wave parameters and statistical parameters in each sub-band signal	Discrete wavelet transform	[33,35–38,44,46,50,52,53,64,76,80,86,88]
	Time-/frequency-domain signals	Use the whole time/frequency domain as the feature	Normalization, windowing, denoising	[29,30,33,45,47,57,60,86–88]
B-scan image	Textual features	Contrast, correlation, energy, entropy, co-occurrence matrix, statistical parameters based on pixel values, etc.	Normalization, noise filtering, segmentation	[74,78,129]
	DWT features	Textural features for each sub-band image	Normalization, DWT	[34,81]
	Image	Use the image as the feature input	Normalization, noise filtering, segmentation	[26,31,58,61,62,90,96,97]

the performance reaches the target, the learning curve method uses a function such as an inverse power law to fit to the data (model performance vs. sample size) [134]. The required sample size may then be determined depending on the target model performance. In contrast to the learning curve method, which attempts to find the required sample size for a target model performance, progressive sampling minimizes the computational cost to reach a target performance, assuming unlimited annotated samples. Besides the empirical methods, model-based approaches have also been proposed to predict the optimal sample size for required model performance [135,139].

5.2.2. Unbalanced data

In addition to determining appropriate data set sizes, good data sets tend to be balanced in that the distribution of signals in each class (crack, non-defect, etc.) is approximately the same. Whereas a slight unbalance across classes is generally not an issue, severe data unbalance can result in inadequate training and less than desired ML model performance. Data unbalances can be caused by biased sampling schemes that result in samples not being collected equally from all classes. The literature has several examples of this, with more ultrasonic measurements collected from cracks and fewer samples from other flaw classes (e.g., LF, LP, PO). Unbalanced data can also be caused by labeling errors (data from one class labeled as another).

Most of the surveyed literature used relatively balanced data sets, though some examples of unbalanced data were reported. In these cases, more measurements were collected from good welds compared to those from welds with incomplete penetration [41], or from cracks compared to porosity [29,30,42,57], or lack of penetration [57]. To

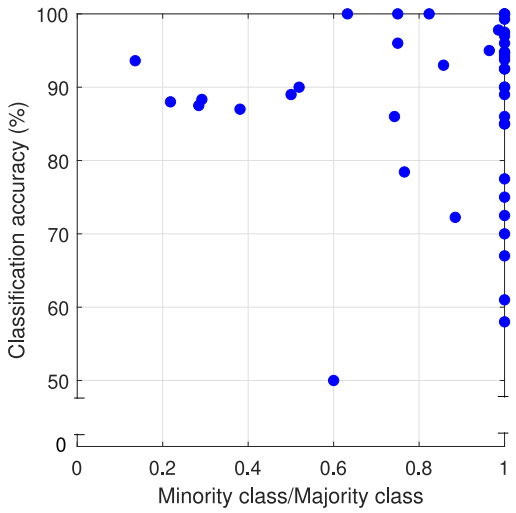

Fig. 9. The ratio of the minority class over majority class vs. classification accuracy.

assess the impact of unbalanced data sets on ML model performance, reported results and the distribution of measurements in the different classes were plotted. Designating the class with the fewest measurements as the minority class and the class with most measurements as the majority class, Fig. 9 shows the reported classification accuracy as a function of the ratio between of the minority and the majority class sizes. Balanced data sets lie at the extreme right of the horizontal axis, corresponding to a ratio close to 1. Although it is not a formal statistical assessment of the impact of unbalanced data sets, the summary in Fig. 9 appears to indicate a generally lower classification accuracy when the data are unbalanced. Perfectly balanced data sets do not necessarily result in greater classification accuracy, indicating that data unbalance is only one element contributing to the overall classification accuracy. One of the challenges with the use of ML for ultrasonic NDE is the potential for small and unbalanced data sets. While there are potentially many options for training with unbalanced data sets and limited data sets, a relatively recent proposal is called an extreme learning machine. While there are some applications of these methods for ultrasonic NDE (for instance, [56,140]), the technique itself appears to need more research on convergence and generalization capabilities.

5.2.3. Data bias

Data bias is a type of error in ML in which certain classes are overweighted, and others are less weighted. A biased dataset will not represent the true dataset, and the learning model trained on the biased dataset will inherit the biases, potentially resulting in inaccurate results. Unbalanced data sets can give rise to data bias, though other causes for data bias can exist. For instance, sample bias arises if the training data are from a single type of sample or a sample with particular characteristics, such as a specific grain structure or surface roughness. Measurement bias occurs when the data collected for the training samples do not represent the data collected in the real world.

Based on the literature surveyed, the extent of the impact of data bias on the reported classification performance for ultrasonic NDE is unclear. Most studies used ultrasonic data from a single type of material (potential sample bias) or a single frequency (potential measurement bias). However, the performance of these ML models on data from other materials or inspection parameters was never tested. It appears that most studies implicitly assumed that these types of biases would lower classification accuracy.

It is worth noting that a few studies trained the ML models using data from two different materials [59,64] in an attempt to increase the diversity of the training data set. Other study [44] used a learning model trained on signals (appropriately scaled) from data of two frequencies (1 MHz and 2.25 MHz), and the results showed that this model had better performance than the model trained on measurements from a single frequency inspection.

Noise could be another source of measurement bias. Most of the studies evaluated herein trained the ML models using low-noise ultrasonic signals. In practice, various sources of noise exist [141], and ML models trained with low-noise data could show poor classification performance if test data have high noise levels. Munir et al. [29] demonstrated this with a DNN. Such studies point to the value of using denoising techniques before training the ML models or using data with different noise levels. Xu et al. [104] demonstrated the performance and denoising capability of clustering algorithm and autoencoder for ultrasound A-scan signals and B-scan images on welding defects. Gantala and Balasubramaniam [62] studied the performance of the same model using training data with and without noises. It demonstrated that the model trained with original data showed better performance than the model trained with noisy data.

Incorrect labeling could also cause measurement bias. This is especially important for supervised learning methods, in which the training data are labeled with the desired class identification. Again, since the literature reviewed did not include information that allowed the evaluation of the extent of this issue or its impact, the authors can only state that studies in the general ML literature indicate incorrect labeling will significantly change the ML classification performance. The training samples for ultrasonic NDE studies should be acquired from verified test sample blocks with known (implanted or grown) flaws types to reduce the bias induced by incorrect labeling. Alternatively, an independent verification technique, such as radiography, should be used where possible (for instance, as performed in other works reviewed [26,53,54,56,97]). Various data validation techniques are recommended in the ML literature for assessing the correctness, completeness, and consistency of the data sets used to train and evaluate ML models, especially when using limited data sets [142–144].

5.3. ML model optimization

ML model parameters (such as neural network weights) are learned from the data and influence the classification performance. The performance can depend on factors external to the model, such as the parameters selected for the training algorithm. These external parameters are called *hyperparameters*, and the selection of optimal values for these hyperparameters is often necessary to optimize classification performance. In the NDE literature, however, this aspect is not usually addressed. Most hyperparameter tuning reported in the reviewed literature used manual selection based on prior experience [34,41,50,99]. Although it is time-consuming, manual tuning of hyperparameters can still improve classification performance. However, this approach may not achieve globally optimal results. Alternative methods for hyperparameter selection exist and have been occasionally applied to the ML for ultrasonic NDE problems. These methods include grid search [32,35], which is suitable for a small number of hyperparameters and random search [145,146], which can be more computationally expensive. Other advanced approaches also exist, such as Bayesian optimization [147,148], Gradient descent [149–153], and genetic algorithms and may provide improved hyperparameter optimization and, consequently, improved classification performance. The results reported in the few articles that employed hyperparameter tuning indicate the value of such optimization. Hyperparameter tuning should be considered a best practice for the use of ML for ultrasonic NDE problems.

5.4. ML model verification and validation (V&V)

ML model V&V is an essential component of quantifying the performance of ML methods and has been widely used to assess the classification performance for ultrasonic NDE. Validation methods often

follow the common practice of designating some of the available data as a test data set. This subset of data is not used in training but is used to evaluate the performance of the trained ML model—essentially acting to verify that the trained ML model does not overfit the training data. Variations in this process include splitting the available data into three subsets – training, validation, and testing – where the validation dataset may be used for hyperparameter tuning [53,56,89,92] or interim checks on model performance during the training process. Cross-validation methods may also be used to optimize the model parameters [35,37,40,50] or to provide a robust estimate of model performance [31,32,41].

These techniques suggest that a separate data set not used as part of the training process should be maintained to yield a robust estimate of the model performance. Part of the challenge with maintaining a test data set is that typically available data sets are small to begin with, as discussed in a previous section. Setting aside data from these data sets can limit the ability of the ML model to learn effectively. A second issue identified in an earlier paragraph is the validation of the data itself [142–144] to ensure that the data used to test the ML models are consistent with the data used to train them, as well as to ensure that the test data are free of errors.

6. Discussion and technical gaps

The literature review indicates that ML methods may be applied to most inspection setups and that there are no inspection-related constraints on the use of ML. The ML methods may likely need to be tuned (model structure, hyperparameters) to maximize performance, but the diversity of data and methods in the literature seems to indicate the potential for widespread use of ML for NDE, including in nuclear power plant in-service inspection.

Other key insights gained to date from the literature analysis include the following.

- Variations in reported classification performance are likely caused by differences in the data and ML algorithms used. Reported data indicate that it is possible to get high true positive rates and low false positive rates simultaneously. However, the best reported results do not appear to be from deep learning algorithms, and it is not clear whether this is an inherent issue with deep learning algorithms or caused by the data sets used in these studies. -
- Reported results indicate no clear correlations between the methods used and classification accuracy. The implication is that most methods are likely capable of good classification performance, with results depending on the data used and model parameter tuning.
- Demonstrating confidence in the results from ML algorithms will require careful attention to data selection, model tuning, and the V&V approach. Representative, common data sets are likely to be necessary to increase confidence in ML performance and allow easier comparison between methods and approaches. V&V approaches to demonstrate the impact of ML on NDE reliability are needed. Methods quantifying confidence bounds or uncertainty in the ML algorithm predictions are also important for assuring the classification results.
- Reproducibility of results requires the publication of the data used, with FAIR (Findability, Accessibility, Interoperability, and Reusability) principles for data stewardship implemented [154]. Availability of common data sets that may be used to develop algorithms and compare results will also enhance the ability to reproduce published results. Publishing a common data set (or multiple data sets) does not preclude setting aside additional data for proprietary purposes or use in blind performance tests (data not included in algorithm training or tuning).
- Data size/representativeness seems to be a limiting condition in most reviewed studies. Data augmentation was used in some cases, though it is not clear whether the approaches increase the information content or just the total number of signals. The actual performance is likely a function of the model parameters, feature selection, and information diversity (not just how many examples are present but the information richness they represent). This observation on data size and representativeness is related to the observations on reference data sets discussed earlier and indicates generalizability of the ML performance must be evaluated using a larger, common data set instead of a subset of study data that may not represent all measurement conditions.
- Sensitivity studies relative to model parameters are likely to be important to improving confidence in the reported results. The impact of noise in the data on the results must be a part of the assessment, and model tuning should be a standard part of the methodology for developing ML solutions.

The previous discussion pointed to limited information in the literature on multiple issues, including data set and sample size determination, dealing with unbalanced data sets and data bias, optimal feature selection, and hyperparameter optimization/model tuning. Reproducibility of results, V&V approaches, confidence/uncertainty estimates, and probability of detection (POD) assessments were other issues with limited information in the literature. The literature also appeared to include limited information on other factors such as software tools, development environment, and staff expertise requirements for proper use and interpretation of these methods. Anecdotal information indicated that software tools and the development environment should be documented for reproducibility. Such documentation would also allow the assessment of potential limitations with these tools.

Many of these topics are actively being explored, and solutions are being proposed in the ML research community. Methods for learning from sparse data sets, propagating uncertainty through ML models, estimating confidence bounds in ML model predictions, and providing limited explainability of ML model performance are discussed in the published literature.

The general ML literature includes methods for data selection, model tuning, sensitivity analyses, V&V, and metrics for performance evaluation. Open-source software for many of these techniques is also available. As a result, it is likely that these techniques, currently primarily applied to the standardized image and time-series data sets, will eventually find their way into the NDE application domain.

7. Conclusions

As computer power increases and the number of qualified NDE inspectors declines, the industry has increased interest in automating some of the analysis of NDE data. As in other science and engineering applications, ML (especially deep learning) is seen as a potential solution to enhancing the reliability of NDE data analysis. Although there has been a significant increase in the number of recent publications in this area (ML for NDE), there appear to be many differences in the way they are applied and great diversity in the methods themselves. A review of the current capabilities of ML and automated data analysis for NDE was conducted to identify any gaps or shortcomings in current ML technology as applied to ultrasonic weld inspections.

The literature review indicated that ML methods may be applied to most encoded ultrasonic inspection setups and that there are no inspection-related constraints on the use of ML. It is likely that the ML method will need to be tuned (model structure, hyperparameters) to maximize performance, but the diversity of data and methods in the literature seems to indicate the potential for widespread use of ML for NDE, including in nuclear power in-service inspection.

Demonstrating confidence in the results from ML algorithms will require careful attention to data selection, model tuning, and the V&V

approach. Representative, common data sets are necessary to increase confidence in ML performance and allow easier comparison between methods and approaches, with V&V strategies needed to demonstrate the impact of ML on NDE reliability. The ability to reproduce the results reported in most studies also appears to be challenging, as it requires access to the data used and details of the algorithms/workflow that were missing in a large number of papers.

The assessed ML for NDE literature appears to have limited information on several issues, including data set and sample size determination, dealing with unbalanced data sets and data bias, optimal feature selection, and hyperparameter optimization/model tuning. Other issues with limited information in the literature included the ability to reproduce results, V&V approaches, confidence/uncertainty estimates, POD, and factors such as software quality assurance (QA) and staff expertise requirements for the proper use and interpretation of these methods. However, these aspects are the target of active research in other applications, and approaches identified in other fields may be applicable to the NDE field.

CRediT authorship contribution statement

Hongbin Sun: Investigation, Formal analysis, Software, Writing – original draft. **Pradeep Ramuhalli:** Conceptualization, Methodology, Writing – review & editing, Project administration, Supervision. **Richard E. Jacob:** Conceptualization, Methodology, Writing – review & editing, Funding acquisition.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the US Nuclear Regulatory Commission (NRC) Office of Research (RES) under Contract 31310019N0001, Task Order 31310020F0038 (Carol Nove, NRC Contracting Officer Representative).

References

[1] B. Bishop, et al., Materials Reliability Program: Risk-Informed Revision of ASME Section XI Appendix G - Proof of Concept (MRP-143), Tech. Rep. 1009546, Electric Power Research Institute, 2005, http://dx.doi.org/10.2172/841931, URL https://www.osti.gov/biblio/841931.

[2] An approach for using probabilistic risk assessment in risk-informed decisions on plant specific changes to the licensing basis, 2011.

[3] L. Udpa, S. Udpa, Eddy current defect characterization using neural networks, NDT Int. 23 (1990) 358.

[4] C.R. Farrar, K. Worden, Structural Health Monitoring: A Machine Learning Perspective, John Wiley & Sons, 2012, http://dx.doi.org/10.1002/9781118443118.

[5] F.-G. Yuan, S.A. Zargar, Q. Chen, S. Wang, Machine learning for structural health monitoring: challenges and opportunities, in: Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2020, Vol. 11379, SPIE, 2020, 1137903, http://dx.doi.org/10.1117/12.2561610.

[6] M. Flah, I. Nunez, W. Ben Chaabene, M.L. Nehdi, Machine learning algorithms in civil structural health monitoring: a systematic review, Arch. Comput. Methods Eng. 28 (4) (2021) 2621–2643, http://dx.doi.org/10.1007/s11831-020-09471-9.

[7] U. Dackermann, B. Skinner, J. Li, Guided wave–based condition assessment of in situ timber utility poles using machine learning algorithms, Struct. Health Monit. 13 (4) (2014) 374–388, http://dx.doi.org/10.1177/1475921714521269.

[8] Y. Liu, Y. Bao, Review on automated condition assessment of pipelines with machine learning, Adv. Eng. Inform. 53 (2022) 101687, http://dx.doi.org/10.1016/j.aei.2022.101687.

[9] ASNT, Nondestructive Testing Handbook, Third Edition: Volume 7, Ultrasonic Testing, American Society for Nondestructive Testing, Columbus, Ohio, 2007.

[10] L. Schmerr, J.-S. Song, Ultrasonic Nondestructive Evaluation Systems: Models and Measurements, first ed., Springer US, 2007, http://dx.doi.org/10.1007/978-0-387-49063-2.

[11] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., Springer, 2017, http://dx.doi.org/10.1007/978-0-387-84858-7.

[12] T. Marwala, Handbook of Machine Learning, in: Handbook of Machine Learning, World Scientific, 2018, http://dx.doi.org/10.1142/11013, URL https://www.worldscientific.com/doi/abs/10.1142/11013.

[13] Encyclopedia of machine learning and data mining, 2017, http://dx.doi.org/10.1007//978-1-4899-7687-1, URL https://rd.springer.com/referencework/10.1007%2F978-1-4899-7687-1.

[14] M.A. Fakih, S. Mustapha, J. Tarraf, G. Ayoub, R. Hamade, Detection and assessment of flaws in friction stir welded joints using ultrasonic guided waves: experimental and finite element analysis, Mech. Syst. Signal Process. 101 (2018) 516–534, http://dx.doi.org/10.1016/j.ymssp.2017.09.003.

[15] X. Yu, P. Zuo, J. Xiao, Z. Fan, Detection of damage in welded joints using high order feature guided ultrasonic waves, Mech. Syst. Signal Process. 126 (2019) 176–192, http://dx.doi.org/10.1016/j.ymssp.2019.02.026.

[16] J.K. Lee, D.S. Bae, S.P. Lee, J.H. Lee, Evaluation on defect in the weld of stainless steel materials using nondestructive technique, Fusion Eng. Des. 89 (7–8) (2014) 1739–1745, http://dx.doi.org/10.1016/j.fusengdes.2013.12.026.

[17] M.G. Droubi, N.H. Faisal, F. Orr, J.A. Steel, M. El-Shaib, Acoustic emission method for defect detection and identification in carbon steel welded joints, J. Construct. Steel Res. 134 (2017) 28–37, http://dx.doi.org/10.1016/j.jcsr.2017.03.012.

[18] T. Sanquist, J. Harrison, Human Factors of Encoded Ultrasonic Examinations in Nuclear Power Plants, Report (NRC ADAMS: ML21124A141), PNNL, 2021.

[19] J. Harrison, R.E. Jacob, M.S. Prowant, A.E. Holmes, C. Hutchinson, A.A. Diaz, Evaluating Flaw Detectability Under Limited-Coverage Conditions, Tech. Rep., PNNL, 2020.

[20] O. Hedden, D. Cowfer, J. Batey, J. Spanner, L. Becker, Overview of the impact of ultrasonic examination performance demonstration on the ASME boiler and pressure vessel code, J. Press. Vessel Technol. 124 (3) (2002) 254–260, http://dx.doi.org/10.1115/1.1490932.

[21] S. Doctor, S.E. Cumblidge1, T.T. Taylor, M.T. Anderson, The Technical Basis Supporting ASME Code, Section XI, Appendix VIII: Performance Demonstration for Ultrasonic Examination Systems, Tech. Rep (NUREG/CR-7165), US NRC, 2013.

[22] J. Welter, J.C. Aldrin, D.S. Forsyth, Automated data analysis (ADA) of ultrasonic NDE data for composites, in: 19th World Conference on Non-Destructive Testing, 2016, pp. 13–17.

[23] J.B. Harley, D. Sparkman, Machine learning and NDE: Past, present, and future, in: AIP Conference Proceedings, Vol. 2102, AIP Publishing LLC, 2019, 090001, http://dx.doi.org/10.1063/1.5099819.

[24] C. Wunderlich, C. Tschöpe, F. Duckhorn, Advanced methods in NDE using machine learning approaches, in: AIP Conference Proceedings, Vol. 1949, AIP Publishing LLC, 2018, 020022, http://dx.doi.org/10.1063/1.5031519.

[25] A.L. Bowler, M.P. Pound, N.J. Watson, A review of ultrasonic sensing and machine learning methods to monitor industrial processes, Ultrasonics 124 (2022) 106776, http://dx.doi.org/10.1016/j.ultras.2022.106776.

[26] I. Virkkunen, T. Koskinen, O. Jessen-Juhler, J. Rinta-Aho, Augmented ultrasonic data for machine learning, J. Nondestruct. Eval. 40 (1) (2021) 1–11, http://dx.doi.org/10.1007/s10921-020-00739-5.

[27] J.F. Lancaster, Metallurgy of Welding, Elsevier, 1999.

[28] A.W. Society, Welding Inspection Handbook, American Welding Society, 2015.

[29] N. Munir, H.-J. Kim, J. Park, S.-J. Song, S.-S. Kang, Convolutional neural network for ultrasonic weldment flaw classification in noisy conditions, Ultrasonics 94 (2019) 74–81, http://dx.doi.org/10.1016/j.ultras.2018.12.001.

[30] N. Munir, J. Park, H.-J. Kim, S.-J. Song, S.-S. Kang, Performance enhancement of convolutional neural network for ultrasonic flaw classification by adopting autoencoder, NDT E Int. 111 (2020) 102218, http://dx.doi.org/10.1016/j.ndteint.2020.102218.

[31] O. Siljama, et al., Reliable Defect Detection Using Machine Learning for Ultrasonic Inspection of Nuclear Power Plant Welds (Master's thesis), Aalto University, 2020.

[32] Y. Yan, D. Liu, B. Gao, G. Tian, Z. Cai, A deep learning-based ultrasonic pattern recognition method for inspecting Girth Weld Cracking of gas pipeline, IEEE Sens. J. 20 (14) (2020) 7997–8006, http://dx.doi.org/10.1109/JSEN.2020.2982680.

[33] E.P. Moura, R.R. Silva, M.H. Siqueira, J.M.A. Rebello, Pattern recognition of weld defects in preprocessed TOFD signals using linear classifiers, J. Nondestruct. Eval. 23 (4) (2004) 163–172, http://dx.doi.org/10.1007/s10921-004-0822-4.

[34] A. Al-Ataby, W. Al-Nuaimy, C. Brett, O. Zahran, Automatic detection and classification of weld flaws in TOFD data using wavelet transform and support vector machines, Insight-Non-Destr. Test. Cond. Monit. 52 (11) (2010) 597–602, http://dx.doi.org/10.1784/insi.2010.52.11.597.

[35] G.A. Guarneri, F.N. Junior, L. de Arruda, Weld discontinuities classification using principal component analysis and support vector machine, in: XI Simpósio Brasileiro de Automaçao Inteligente, 2013, pp. 2358–4483.

[36] Y. Chen, H.-W. Ma, G.-M. Zhang, A support vector machine approach for classification of welding defects from ultrasonic signals, Nondestr. Test. Eval. 29 (3) (2014) 243–254, http://dx.doi.org/10.1080/10589759.2014.914210.

[37] K. Virupakshappa, E. Oruklu, Ultrasonic flaw detection using support vector machine classification, in: 2015 IEEE International Ultrasonics Symposium (IUS), IEEE, Taipei, Taiwan, 2015, pp. 1–4, http://dx.doi.org/10.1109/ULTSYM.2015.0128.

[38] X. Wang, S. Guan, L. Hua, B. Wang, X. He, Classification of spot-welded joint strength using ultrasonic signal time-frequency features and PSO-SVM method, Ultrasonics 91 (2019) 161–169, http://dx.doi.org/10.1016/j.ultras.2018.08.014.

[39] H. Xiao, D. Chen, J. Xu, S. Guo, Defects identification using the improved ultrasonic measurement model and support vector machines, NDT E Int. 111 (2020) 102223, http://dx.doi.org/10.1016/j.ndteint.2020.102223.

[40] T.M. Nunes, V.H.C. De Albuquerque, J.P. Papa, C.C. Silva, P.G. Normando, E.P. Moura, J.M.R. Tavares, Automatic microstructural characterization and classification using artificial intelligence techniques on ultrasound signals, Expert Syst. Appl. 40 (8) (2013) 3096–3105, http://dx.doi.org/10.1016/j.eswa.2012.12.025.

[41] Ó. Martín, M. Pereda, J.I. Santos, J.M. Galán, Assessment of resistance spot welding quality based on ultrasonic testing and tree-based techniques, J. Mater Process. Technol. 214 (11) (2014) 2478–2487, http://dx.doi.org/10.1016/j.jmatprotec.2014.05.021.

[42] S. Song, L. Schmerr, Ultrasonic flaw classification in weldments using probabilistic neural networks, J. Nondestruct. Eval. 11 (2) (1992) 69–77, http://dx.doi.org/10.1007/BF00568290.

[43] A. Masnata, M. Sunseri, Neural network classification of flaws detected by ultrasonic means, NDT E Int. 29 (2) (1996) 87–93, http://dx.doi.org/10.1016/0963-8695(95)00053-4.

[44] R. Polikar, L. Udpa, S.S. Udpa, T. Taylor, Frequency invariant classification of ultrasonic weld inspection signals, IEEE Trans. Ultrason. Ferroelectr. Freq. Control 45 (3) (1998) 614–625, http://dx.doi.org/10.1109/58.677606.

[45] F. Margrave, K. Rigas, D. Bradley, P. Barrowcliffe, The use of neural networks in ultrasonic flaw detection, Measurement 25 (2) (1999) 143–154, http://dx.doi.org/10.1016/S0263-2241(98)00075-X.

[46] S. Legendre, D. Massicotte, J. Goyette, T. Bose, Neural classification of lamb wave ultrasonic weld testing signals using wavelet coefficients, IEEE Trans. Instrum. Meas. 50 (3) (2001) http://dx.doi.org/10.1109/19.930439, Conference Name: IEEE Transactions on Instrumentation and Measurement.

[47] J. Veiga, A. De Carvalho, I. Da Silva, J. Rebello, The use of artificial neural network in the classification of pulse-echo and TOFD ultra-sonic signals, J. Braz. Soc. Mech. Sci. Eng. 27 (4) (2005) 394–398, http://dx.doi.org/10.1590/S1678-58782005000400007.

[48] A. Carvalho, J. Rebello, L. Sagrilo, C. Camerini, I. Miranda, MFL signals and artificial neural networks applied to detection and classification of pipe weld defects, Ndt E Int. 39 (8) (2006) 661–667, http://dx.doi.org/10.1016/j.ndteint.2006.04.003.

[49] Ó. Martín, M. López, F. Martín, Artificial neural networks for quality control by ultrasonic testing in resistance spot welding, J. Mater Process. Technol. 183 (2–3) (2007) 226–233, http://dx.doi.org/10.1016/j.jmatprotec.2006.10.011.

[50] S. Sambath, P. Nagaraj, N. Selvakumar, Automatic defect classification in ultrasonic NDT using artificial intelligence, J. Nondestruct. Eval. 30 (1) (2011) 20–28, http://dx.doi.org/10.1007/s10921-010-0086-0.

[51] S. Lalithakumari, B. Sheelarani, B. Venkatraman, Artificial neural network based defect detection of welds in TOFD technique, Int. J. Comput. Appl. 41 (20) (2012) http://dx.doi.org/10.5120/5808-8069.

[52] S. Seyedtabaii, Performance evaluation of neural network based pulse-echo weld defect classifiers, Meas. Sci. Rev. 12 (5) (2012) 168–174, http://dx.doi.org/10.2478/v10048-012-0026-5.

[53] F. Cruz, E. Simas Filho, M. Albuquerque, I. Silva, C. Farias, L. Gouvêa, Efficient feature selection for neural network based detection of flaws in steel welded joints using ultrasound testing, Ultrasonics 73 (2017) 1–8, http://dx.doi.org/10.1016/j.ultras.2016.08.017.

[54] J. Liu, G. Xu, L. Ren, Z. Qian, L. Ren, Defect intelligent identification in resistance spot welding ultrasonic detection based on wavelet packet and neural network, Int. J. Adv. Manuf. Technol. 90 (9–12) (2017) 2581–2588, http://dx.doi.org/10.1007/s00170-016-9588-y.

[55] S. Lalithakumari, R. Pandian, Effect of topology changes of neural network in classification of weld defects, Mater. Today: Proc. 33 (2020) 2656–2659, http://dx.doi.org/10.1016/j.matpr.2020.01.222.

[56] L.C. Silva, E.F. Simas Filho, M.C. Albuquerque, I.C. Silva, C.T. Farias, Segmented analysis of time-of-flight diffraction ultrasound for flaw detection in welded steel plates using extreme learning machines, Ultrasonics 102 (2020) 106057, http://dx.doi.org/10.1016/j.ultras.2019.106057.

[57] N. Munir, H.-J. Kim, S.-J. Song, S.-S. Kang, Investigation of deep neural network with drop out for ultrasonic flaw classification in weldments, J. Mech. Sci. Technol. 32 (7) (2018) 3073–3080, http://dx.doi.org/10.1007/s12206-018-0610-1.

[58] H. Zhu, W. Ge, Z. Liu, Deep learning-based classification of weld surface defects, Appl. Sci. 9 (16) (2019) 3312, http://dx.doi.org/10.3390/app9163312.

[59] K. Sudheera, N. Nandhitha, V.B.V. Sai, N.V. Kumar, Deep learning techniques for flaw characterization in weld pieces from ultrasonic signals, Russ. J. Nondestr. Test. 56 (10) (2020) 820–830, http://dx.doi.org/10.1134/S1061830920100083.

[60] J. Park, S.-E. Lee, H.-J. Kim, S.-J. Song, S.-S. Kang, System invariant method for ultrasonic flaw classification in weldments using residual neural network, Appl. Sci. 12 (3) (2022) http://dx.doi.org/10.3390/app12031477, URL https://www.mdpi.com/2076-3417/12/3/1477.

[61] T. Latête, B. Gauthier, P. Belanger, Towards using convolutional neural network to locate, identify and size defects in phased array ultrasonic testing, Ultrasonics 115 (2021) 106436, http://dx.doi.org/10.1016/j.ultras.2021.106436.

[62] T. Gantala, K. Balasubramaniam, Automated defect recognition for welds using simulation assisted tfm imaging with artificial intelligence, J. Nondestruct. Eval. 40 (1) (2021) 1–24, http://dx.doi.org/10.1007/s10921-021-00761-1.

[63] E. De Moura, M. Siqueira, R. da Silva, J. Rebello, L. Calôba, Welding defect pattern recognition in TOFD signals Part 1. Linear classifiers, Insight-Non-Destr. Test. Cond. Monit. 47 (12) (2005) 777–782, http://dx.doi.org/10.1784/insi.2005.47.12.777.

[64] V. Matz, M. Kreidl, R. Smid, Classification of ultrasonic signals, Int. J. Mater. Product Technol. 27 (3–4) (2006) 145–155, http://dx.doi.org/10.1504/IJMPT.2006.011267.

[65] N.A. Akram, D. Isa, R. Rajkumar, L.H. Lee, Active incremental support vector machine for oil and gas pipeline defects prediction system using long range ultrasonic transducers, Ultrasonics 54 (6) (2014) 1534–1544, http://dx.doi.org/10.1016/j.ultras.2014.03.017.

[66] Y. Yuan, K. Virupakshappa, Y. Jiang, E. Oruklu, Comparison of GPU and FPGA based hardware platforms for ultrasonic flaw detection using support vector machines, in: 2017 IEEE International Ultrasonics Symposium (IUS), (ISSN: 1948-5727) 2017, pp. 1–4, http://dx.doi.org/10.1109/ULTSYM.2017.8092499.

[67] Y. Chen, H.-W. Ma, M. Dong, Automatic classification of welding defects from ultrasonic signals using an SVM-based RBF neural network approach, Insight - Non-Destr. Test. Cond. Monit. 60 (4) (2018) 194–199, http://dx.doi.org/10.1784/insi.2018.60.4.194.

[68] J.-G. Kim, C. Jang, S.-S. Kang, Classification of ultrasonic signals of thermally aged cast austenitic stainless steel (CASS) using machine learning (ML) models, Nucl. Eng. Technol. (2021) http://dx.doi.org/10.1016/j.net.2021.09.033, URL https://www.sciencedirect.com/science/article/pii/S173857332100574X.

[69] R. Polikar, L. Udpa, S. Udpa, V. Honavar, An incremental learning algorithm with confidence estimation for automated identification of NDE signals, IEEE Trans. Ultrason. Ferroelectr. Freq. Control 51 (8) (2004) 990–1001, http://dx.doi.org/10.1109/TUFFC.2004.1324403.

[70] P. Madhumitha, S. Ramkishore, K. Srikanth, P. Palanichamy, Application of decision trees for the identification of weld central line in austenitic stainless steel weld joints, in: 2014 International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC), IEEE, 2014, pp. 400–406, http://dx.doi.org/10.1109/ICCPEIC.2014.6915397.

[71] N. Shipway, T. Barden, P. Huthwaite, M. Lowe, Automated defect detection for fluorescent penetrant inspection using random forest, NDT E Int. 101 (2019) 113–123, http://dx.doi.org/10.1016/j.ndteint.2018.10.008.

[72] N. Shipway, P. Huthwaite, M. Lowe, T. Barden, Performance based modifications of random forest to perform automated defect detection for fluorescent penetrant inspection, J. Nondestruct. Eval. 38 (2) (2019) 1–11, http://dx.doi.org/10.1007/s10921-019-0574-9.

[73] Z. Zhang, Z. Yang, W. Ren, G. Wen, Random forest-based real-time defect detection of Al alloy in robotic arc welding using optical spectrum, J. Manuf. Process. 42 (2019) 51–59, http://dx.doi.org/10.1016/j.jmapro.2019.04.023.

[74] S.W. Lawson, G.A. Parker, Automatic detection of defects in industrial ultrasound images using a neural network, in: Vision Systems: Applications, Vol. 2786, International Society for Optics and Photonics, 1996, pp. 37–47, http://dx.doi.org/10.1117/12.248579.

[75] S.-J. Song, H.-J. Kim, H. Cho, Development of an intelligent system for ultrasonic flaw classification in weldments, Nucl. Eng. Des. 212 (1) (2002) 307–320, http://dx.doi.org/10.1016/S0029-5493(01)00495-2.

[76] F. Bettayeb, T. Rachedi, H. Benbartaoui, An improved automated ultrasonic NDE system by wavelet and neuron networks, Ultrasonics 42 (1) (2004) 853–858, http://dx.doi.org/10.1016/j.ultras.2004.01.064, URL https://www.sciencedirect.com/science/article/pii/S0041624X04000721.

[77] E. de Moura, M. Siqueira, R. da Silva, J. Rebello, Welding defect pattern recognition in TOFD signals Part 2. Non-linear classifiers, Insight-Non-Destr. Test. Cond. Monit. 47 (12) (2005) 783–787, http://dx.doi.org/10.1784/insi.2005.47.12.783.

[78] C.N. Shitole, O. Zahran, W. Al-Nuaimy, Combining fuzzy logic and neural networks in classification of weld defects using ultrasonic time-of-flight diffraction, Insight-Non-Destr. Test. Cond. Monit. 49 (2) (2007) 79–82, http://dx.doi.org/10.1784/insi.2007.49.2.79.

[79] S. Lalithakumari, B. Bseelarani, B. Venkatraman, Classification of TOFD signals by artificial neural network, in: 18th World Conference on Nondestructive Testing, 2012, pp. 1–6.

[80] L. Yang, I.C. Ume, Measurement of weld penetration depths in thin structures using transmission coefficients of laser-generated lamb waves and neural network, Ultrasonics 78 (2017) 96–109, http://dx.doi.org/10.1016/j.ultras.2017.02.019.

[81] F. Roca Barceló, P. Jaén del Hierro, F. Ribes Llario, J. Real Herráiz, Development of an ultrasonic weld inspection system based on image processing and neural networks, Nondestr. Test. Eval. 33 (2) (2018) 229–236, http://dx.doi.org/10.1080/10589759.2017.1376056.

[82] L.F. Rodrigues, F.C. Cruz, M.A. Oliveira, E.F. Simas Filho, M.C. Albuquerque, I.C. Silva, C.T. Farias, Carburization level identification in industrial HP pipes using ultrasonic evaluation and machine learning, Ultrasonics 94 (2019) 145–151, http://dx.doi.org/10.1016/j.ultras.2018.10.005.

[83] Z. Chen, G. Huang, C. Lu, G. Chen, Automatic recognition of weld defects in TOFD D-scan images based on faster R-CNN, J. Test. Eval. 48 (2) (2020) 811–824, http://dx.doi.org/10.1520/JTE20170563.

[84] S.E. Florence, R.V. Samsingh, V. Babureddy, Artificial intelligence based defect classification for weld joints, in: IOP Conference Series: Materials Science and Engineering, Vol. 402, IOP Publishing, 2018, 012159, http://dx.doi.org/10.1088/1757-899X/402/1/012159.

[85] N. Amiri, G. Farrahi, K.R. Kashyzadeh, M. Chizari, Applications of ultrasonic testing and machine learning methods to predict the static & fatigue behavior of spot-welded joints, J. Manuf. Process. 52 (2020) 26–34, http://dx.doi.org/10.1016/j.jmapro.2020.01.047.

[86] P. Pawar, R. Buktar, Detection and classification of defects in ultrasonic testing using deep learning, in: Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications, Springer, 2022, pp. 1–15, http://dx.doi.org/10.1007/978-981-16-6407-6_1.

[87] F. Bettayeb, H. Benbartaoui, B. Raouraou, The reliability of the ultrasonic characterization of welds by the artificial neural network, in: 17th World Conference on Nondestructive Testing, 2008, pp. 25–28.

[88] I.S. Souza, M.C. Albuquerque, E.F. de SIMAS FILHO, C.T. FARIAS, Signal processing techniques for ultrasound automatic identification of flaws in steel welded joints–a comparative analysis, in: 18th World Conference on Nondestructive Testing, 2012, pp. 16–20.

[89] K. Virupakshappa, E. Oruklu, Multi-class classification of defect types in ultrasonic NDT signals with convolutional neural networks, in: 2019 IEEE International Ultrasonics Symposium (IUS), IEEE, 2019, pp. 1647–1650, http://dx.doi.org/10.1109/ULTSYM.2019.8926027.

[90] Z.N. Reza, Real-Time Automated Weld Quality Analysis from Ultrasonic B-Scan Using Deep Learning (Ph.D. thesis), University of Windsor (Canada), 2019.

[91] S. Sudhagar, M. Sakthivel, P. Ganeshkumar, Monitoring of friction stir welding based on vision system coupled with machine learning algorithm, Measurement 144 (2019) 135–143, http://dx.doi.org/10.1016/j.measurement.2019.05.018.

[92] Q. Wang, W. Jiao, P. Wang, Y. Zhang, A tutorial on deep learning-based data analytics in manufacturing through a welding case study, J. Manuf. Process. 63 (2021) 2–13, http://dx.doi.org/10.1016/j.jmapro.2020.04.044.

[93] F. Gao, B. Li, L. Chen, X. Wei, Z. Shang, C. He, Ultrasonic signal denoising based on autoencoder, Rev. Sci. Instrum. 91 (4) (2020) 045104, http://dx.doi.org/10.1063/1.5136269.

[94] M.D. Jedrusiak, A deep learning approach for denoising air-coupled ultrasonic responds data, Int. J. Artif. Intell. Appl. (IJAIA) 11 (4) (2020) http://dx.doi.org/10.5121/ijaia.2020.11402.

[95] I. Virkkunen, T. Koskinen, Flaw detection in ultrasonic data using deep learning, in: Baltica XI 2019: International Conference on Life Management and Maintenance for Power Plants, VTT Technical Research Centre of Finland, 2019, pp. 1–8.

[96] E. Provencal, L. Laperrière, Identification of weld geometry from ultrasound scan data using deep learning, Proc. CIRP 104 (2021) 122–127, http://dx.doi.org/10.1016/j.procir.2021.11.021.

[97] T. Koskinen, I. Virkkunen, O. Siljama, O. Jessen-Juhler, The effect of different flaw data to machine learning powered ultrasonic inspection, J. Nondestruct. Eval. 40 (1) (2021) 1–13, http://dx.doi.org/10.1007/s10921-021-00757-x.

[98] R. Otero, C. Correia, C. Ruiz, J. Michinaux, L. Bravo, Statistical characterization from ultrasonic signals using time-frequency representation, J. Nondestr. Test. 8 (5) (2003) 1–10.

[99] R.H. Murta, F.d.A. Vieira, V.O. Santos, E.P. de Moura, Welding defect classification from simulated ultrasonic signals, J. Nondestruct. Eval. 37 (3) (2018) 40, http://dx.doi.org/10.1007/s10921-018-0496-y.

[100] K. Virupakshappa, E. Oruklu, Unsupervised machine learning for ultrasonic flaw detection using Gaussian mixture modeling, K-means clustering and mean shift clustering, in: 2019 IEEE International Ultrasonics Symposium (IUS), IEEE, 2019, pp. 647–649, http://dx.doi.org/10.1109/ULTSYM.2019.8926078.

[101] B. Cassels, L.-K. Shark, S.J. Mein, A. Nixon, T. Barber, R. Turner, Robust principal component analysis of ultrasonic sectorial scans for defect detection in weld inspection, in: Multimodal Sensing: Technologies and Applications, Vol. 11059, International Society for Optics and Photonics, 2019, 110590E, http://dx.doi.org/10.1117/12.2527622.

[102] A.L. Qi, J.F. Wang, F. Wang, U. Idachaba, G. Akanmu, Welding defect classification of ultrasonic detection based on PCA and KNN, in: Vehicle, Mechatronics and Information Technologies, in: Applied Mechanics and Materials, vol. 380, Trans Tech Publications Ltd, 2013, pp. 902–906, http://dx.doi.org/10.4028/www.scientific.net/AMM.380-384.902.

[103] L.A. Martins, F.L. Pádua, P.E. Almeida, Automatic detection of surface defects on rolled steel using computer vision and artificial neural networks, in: IECON 2010-36th Annual Conference on IEEE Industrial Electronics Society, IEEE, 2010, pp. 1081–1086, http://dx.doi.org/10.1109/IECON.2010.5675519.

[104] W. Xu, X. Li, J. Zhang, Z. Xue, J. Cao, Ultrasonic signal enhancement for coarse grain materials by machine learning analysis, Ultrasonics 117 (2021) 106550, http://dx.doi.org/10.1016/j.ultras.2021.106550, URL https://linkinghub.elsevier.com/retrieve/pii/S0041624X21001773.

[105] S.A. Shevchik, T. Le-Quang, F.V. Farahani, N. Faivre, B. Meylan, S. Zanoli, K. Wasmer, Laser welding quality monitoring via graph support vector machine with data adaptive kernel, IEEE Access 7 (2019) 93108–93122, http://dx.doi.org/10.1109/ACCESS.2019.2927661, Conference Name: IEEE Access.

[106] T.G. Dietterich, Ensemble methods in machine learning, in: International Workshop on Multiple Classifier Systems, Springer, 2000, pp. 1–15, http://dx.doi.org/10.1007/3-540-45014-9_1.

[107] C. Zhang, Y. Ma, Ensemble Machine Learning: Methods and Applications, Springer, 2012, http://dx.doi.org/10.1007/978-1-4419-9326-7.

[108] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, Y. Zhou, A novel ensemble method for classifying imbalanced data, Pattern Recognit. 48 (5) (2015) 1623–1637, http://dx.doi.org/10.1016/j.patcog.2014.11.014.

[109] D. Parikh, M.T. Kim, J. Oagaro, S. Mandayam, R. Polikar, Ensemble of classifiers approach for NDT data fusion, in: IEEE Ultrasonics Symposium, 2004, Vol. 2, IEEE, 2004, pp. 1062–1065, http://dx.doi.org/10.1109/ULTSYM.2004.1417959.

[110] J. Camacho-Navarro, M. Ruiz, R. Villamizar, L. Mujica, G. Moreno-Beltrán, Ensemble learning as approach for pipeline condition assessment, in: Journal of Physics: Conference Series, Vol. 842, IOP Publishing, 2017, 012019, http://dx.doi.org/10.1088/1742-6596/842/1/012019.

[111] H. Baumgartl, J. Tomas, R. Buettner, M. Merkel, A novel deep-learning approach for automated non-destructive testing in quality assurance based on convolutional neural networks, in: ACEX-2019 Proceedings, 2019.

[112] D. Mery, C. Arteta, Automatic defect recognition in x-ray testing using computer vision, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2017, pp. 1026–1035, http://dx.doi.org/10.1109/WACV.2017.119.

[113] Q. Luo, B. Gao, W.L. Woo, Y. Yang, Temporal and spatial deep learning network for infrared thermal defect detection, NDT E Int. 108 (2019) 102164, http://dx.doi.org/10.1016/j.ndteint.2019.102164.

[114] Y. Duan, S. Liu, C. Hu, J. Hu, H. Zhang, Y. Yan, N. Tao, C. Zhang, X. Maldague, Q. Fang, et al., Automated defect classification in infrared thermography based on a neural network, NDT E Int. 107 (2019) 102147, http://dx.doi.org/10.1016/j.ndteint.2019.102147.

[115] A. Géron, Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, O'Reilly Media, 2019.

[116] W. Zhang, C. Li, G. Peng, Y. Chen, Z. Zhang, A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load, Mech. Syst. Signal Process. 100 (2018) 439–453, http://dx.doi.org/10.1016/j.ymssp.2017.06.022.

[117] D.-A. Clevert, T. Unterthiner, S. Hochreiter, Fast and accurate deep network learning by exponential linear units (elus), 2015, arXiv preprint arXiv:1511.07289.

[118] X.-X. Niu, C.Y. Suen, A novel hybrid CNN–SVM classifier for recognizing handwritten digits, Pattern Recognit. 45 (4) (2012) 1318–1325, http://dx.doi.org/10.1016/j.patcog.2011.09.021.

[119] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65, http://dx.doi.org/10.1016/0377-0427(87)90125-7.

[120] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. 2 (1979) 224–227, http://dx.doi.org/10.1109/TPAMI.1979.4766909.

[121] A. Ortiz, F. Bonnin-Pascual, E. Garcia-Fidalgo, J.P. Company-Corcoles, Vision-based corrosion detection assisted by a micro-aerial vehicle in a vessel inspection application, Sensors 16 (12) (2016) 2118, http://dx.doi.org/10.3390/s16122118.

[122] B.T. Bastian, J. N, S.K. Ranjith, C.V. Jiji, Visual inspection and characterization of external corrosion in pipelines using deep neural network, NDT E Int. 107 (2019) 102134, http://dx.doi.org/10.1016/j.ndteint.2019.102134.

[123] F. Chen, M.R. Jahanshahi, NB-CNN: Deep learning-based crack detection using convolutional neural network and naïve Bayes data fusion, IEEE Trans. Ind. Electron. 65 (5) (2018) 4392–4400, http://dx.doi.org/10.1117/12.2296772, Conference Name: IEEE Transactions on Industrial Electronics.

[124] T. Papamarkou, H. Guy, B. Kroencke, J. Miller, P. Robinette, D. Schultz, J. Hinkle, L. Pullum, C. Schuman, J. Renshaw, S. Chatzidakis, Automated detection of corrosion in used nuclear fuel dry storage canisters using residual neural networks, 2020, arXiv:2003.03241 [Cs, Stat].

[125] J.-x. Duan, L. Luo, X.-r. Gao, J.-p. Peng, J.-l. Li, Hybrid ultrasonic TOFD imaging of weld flaws using wavelet transforms and image registration, J. Nondestruct. Eval. 37 (2) (2018) 1–11, http://dx.doi.org/10.1007/s10921-018-0476-2.

[126] F. Yeh, T. Lukomski, J. Haag, T. Clarke, T. Stepinski, T. Strohaecker, An alternative ultrasonic TimeofFlight diffraction (TOFD) method, Ndt E Int. 100 (2018) 74–83, http://dx.doi.org/10.1016/j.ndteint.2018.08.008.

[127] S. Bae, L. Udpa, S. Udpa, T. Taylor, Classification of ultrasonic weld inspection data using principal component analysis, in: Review of Progress in Quantitative Nondestructive Evaluation, Springer, 1997, pp. 741–748, http://dx.doi.org/10.1007/978-1-4615-5947-4_97.

[128] K. Sudheera, N. Nandhitha, P. Nanekar, B. Venkatraman, B.S. Rani, Automated weld defect classification from ultrasonic signals using statistical moments on normal distribution curves of wavelet co-efficient, in: 2013 International Conference on Advanced Electronic Systems (ICAES), IEEE, 2013, pp. 24–28, http://dx.doi.org/10.1109/ICAES.2013.6659354.

[129] O. Zahran, W. Nuaimy, Automatic classification of defects in time-of-flight diffraction data, Ndt.Net 30 (2004).

[130] J. Cai, J. Luo, S. Wang, S. Yang, Feature selection in machine learning: A new perspective, Neurocomputing 300 (2018) 70–79, http://dx.doi.org/10.1016/j.neucom.2017.11.077.

[131] P.M. Granitto, C. Furlanello, F. Biasioli, F. Gasperi, Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products, Chemometr. Intell. Lab. Syst. 83 (2) (2006) 83–90, http://dx.doi.org/10.1016/j.chemolab.2006.01.007.

[132] J. Reunanen, Overfitting in making comparisons between variable selection methods, J. Mach. Learn. Res. 3 (Mar) (2003) 1371–1382.

[133] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electr. Eng. 40 (1) (2014) 16–28, http://dx.doi.org/10.1016/j.compeleceng.2013.11.024.

[134] S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell, T.R. Golub, J.P. Mesirov, Estimating dataset size requirements for classifying DNA microarray data, J. Comput. Biol. 10 (2) (2003) 119–142, http://dx.doi.org/10.1089/106652703321825928.

[135] K.K. Dobbin, Y. Zhao, R.M. Simon, How large a training set is needed to develop a classifier for microarray data? Clin. Cancer Res. 14 (1) (2008) 108–114, http://dx.doi.org/10.1158/1078-0432.CCR-07-0443.

[136] C. Meek, B. Thiesson, D. Heckerman, The learning-curve sampling method applied to model-based clustering, J. Mach. Learn. Res. 2 (Feb) (2002) 397–418.

[137] C. Perlich, F. Provost, J.S. Simonoff, Tree induction vs. logistic regression: A learning-curve analysis, J. Mach. Learn. Res. 4 (Jun) (2003) 211–255.

[138] R.L. Figueroa, Q. Zeng-Treitler, S. Kandula, L.H. Ngo, Predicting sample size required for classification performance, BMC Med. Inform. Decis. Mak. 12 (1) (2012) 8, http://dx.doi.org/10.1186/1472-6947-12-8.

[139] R.H. Byrd, G.M. Chin, J. Nocedal, Y. Wu, Sample size selection in optimization methods for machine learning, Math. Program. 134 (1) (2012) 127–155, http://dx.doi.org/10.1007/s10107-012-0572-5.

[140] L.C. Silva, E.F. Simas Filho, M.C. Albuquerque, I.C. Silva, C.T. Farias, Embedded decision support system for ultrasound nondestructive evaluation based on extreme learning machines, Comput. Electr. Eng. 90 (2021) 106891, http://dx.doi.org/10.1016/j.compeleceng.2020.106891.

[141] B. Smagowska, Ultrasonic noise sources in a work environment, Arch. Acoust. 38 (2) (2013) 169–176, http://dx.doi.org/10.1121/1.5063812.

[142] F. Biessmann, J. Golebiowski, T. Rukat, D. Lange, P. Schmidt, Automated data validation in machine learning systems, Bull. IEEE Comput. Soc. Tech. Comm. Data Eng. (2015).

[143] H. Foidl, M. Felderer, Risk-based data validation in machine learning-based software systems, in: MaLTeSQuE 2019: Proceedings of the 3rd ACM SIGSOFT International Workshop on Machine Learning Techniques for Software Quality Evaluation, Association for Computing Machinery, New York, NY, USA, 2019, pp. 13–18, http://dx.doi.org/10.1145/3340482.3342743.

[144] A. Vabalas, E. Gowen, E. Poliakoff, A.J. Casson, Machine learning algorithm validation with a limited sample size, PLoS One 14 (11) (2019) e0224365, http://dx.doi.org/10.1371/journal.pone.0224365.

[145] J. Bergstra, R. Bardenet, Y. Bengio, B. Kégl, Algorithms for hyper-parameter optimization, in: 25th Annual Conference on Neural Information Processing Systems (NIPS 2011), Vol. 24, Neural Information Processing Systems Foundation, 2011, pp. 2546–2554.

[146] J. Bergstra, Y. Bengio, Random search for hyperparameter optimization, J. Mach. Learn. Res. 13 (2) (2012) 281–305.

[147] J. Snoek, H. Larochelle, R.P. Adams, Practical bayesian optimization of machine learning algorithms, 2012, arXiv preprint arXiv:1206.2944.

[148] A. Klein, S. Falkner, S. Bartels, P. Hennig, F. Hutter, Fast bayesian optimization of machine learning hyperparameters on large datasets, in: Artificial Intelligence and Statistics, PMLR, 2017, pp. 528–536.

[149] D. Maclaurin, D. Duvenaud, R. Adams, Gradient-based hyperparameter optimization through reversible learning, in: International Conference on Machine Learning, PMLR, 2015, pp. 2113–2122.

[150] J. Domke, Generic methods for optimization-based modeling, in: Artificial Intelligence and Statistics, PMLR, 2012, pp. 318–326.

[151] J. Luketina, M. Berglund, K. Greff, T. Raiko, Scalable gradient-based tuning of continuous regularization hyperparameters, in: International Conference on Machine Learning, PMLR, 2016, pp. 2952–2960.

[152] F. Pedregosa, Hyperparameter optimization with approximate gradient, in: International Conference on Machine Learning, PMLR, 2016, pp. 737–746.

[153] J. Fu, H. Luo, J. Feng, K.H. Low, T.-S. Chua, Drmad: distilling reverse-mode automatic differentiation for optimizing hyperparameters of deep neural networks, 2016, http://dx.doi.org/10.48550/arXiv.1601.00917, arXiv preprint arXiv:1601.00917.

[154] M. Wilkinson, M. Dumontier, I. Aalbersberg, et al., The FAIR guiding principles for scientific data management and stewardship, Sci. Data 3 (2016) 160018, http://dx.doi.org/10.1038/sdata.2016.18.