ORNL/SPR-2023/3245

An Assessment of Machine Learning Applied to Ultrasonic Nondestructive Evaluation

Hongbin Sun
Pradeep Ramuhalli
Ryan Meyer

December 2023

OAK RIDGE National Laboratory

ORNL IS MANAGED BY UT-BATTELLE LLC FOR THE US DEPARTMENT OF ENERGY

ORNL/SPR-2023/3245

Nuclear Energy and Fuel Cycle Division

AN ASSESSMENT OF MACHINE LEARNING APPLIED TO ULTRASONIC NONDESTRUCTIVE EVALUATION

Hongbin Sun
Pradeep Ramuhalli
Ryan Meyer

December 2023

Prepared for the U.S. Nuclear Regulatory Commission
Office of Nuclear Regulatory Research
Under Contract DE-AC05-00OR22725
Interagency Agreement: 31310022S0022
Carol A. Nove – NRC Contracting Officer Representative

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AI	Artificial Intelligence
ASME	American Society of Mechanical Engineers
BPVC	(ASME) Boiler and Pressure Vessel Code
CNN	Convolutional Neural Network
CS	Condition Assessment
DL	Deep learning
DMW	Dissimilar Material Weld
DNN	Deep Neural Network
DOE	U.S. Department of Energy
EDM	Electrical Discharge Machined
FAIR	Findability, Accessibility, Interoperability, and Reusability
FN	False Negative
FP	False Positive
FPR	False Positive Rate
HAZ	Heat-Affected Zone
ISI	Inservice Inspection
ML	Machine Learning
NDE	Nondestructive Examination
NDT&E	Nondestructive Testing and Evaluation
NPP	Nuclear Power Plant
OSTI	Office of Scientific and Technical Information
POD	Probability of Detection
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
SHM	Structural Health Monitoring
SS	Stainless Steel
SC	Saw Cut
TFC	Thermal Fatigue Crack
TN	True Negative
TP	True Positive
TPR	True Positive Rate
UT	Ultrasonic testing
V&V	Verification and Validation

ABSTRACT

In the United States, the nuclear industry performs inservice inspection (ISI) through nondestructive examination (NDE) methods in accordance with guidelines specified in the American Society of Mechanical Engineers (ASME) Boiler and Pressure Vessel Code (BPVC), Section XI, Rules for Inservice Inspection of Nuclear Power Plant Components. Ultrasonic nondestructive testing and evaluation (NDT&E) is one of the more commonly used techniques for inspecting Class 1 structural components in nuclear power systems. As the number of qualified NDE inspectors declines, the nuclear industry is looking to take advantage of advances in automation to enhance inspection capabilities. Advances in computational power, cloud-based computing, and machine learning algorithms make automated data analysis possible. Machine learning (ML) has shown huge potential in automated data analyses for ultrasonic NDE in the context of weld inspections.

Questions around the capabilities of ML models for automated data analysis of NDE data and factors that impact ML performance for ultrasonic NDE of welding defects remain. This study included an empirical assessment of ML for ultrasonic NDE, with the objective of identifying factors influencing ML performance. Part of this assessment included the development of a reference dataset consisting of data collected from multiple specimens and flaws with various inspection procedures. Given the focus of the study on the classification of B-scan data into flaw and non-flaw categories, B-scan images were extracted from the data files and compiled into a retrievable common dataset with labels (flaw and non-flaw) as appropriate for each B-scan image. A portion of this dataset served as a reference test data set for evaluating the ML performance analysis. A convolutional neural network (CNN) model was used as the prototypic algorithm for the classification of ultrasonic B-scan images of cracks within weldments in the reference dataset. A subset of the ultrasonic data, composed of data from four stainless-steel weld specimens and two dissimilar metal weld (DMW) specimens with two types of defects [saw cuts and thermal fatigue cracks (TFC)], was used in this initial phase of the assessment. Data from conventional inspections with longitudinal and shear wave transducers collected at multiple inspection frequencies and inspection angles were used in this assessment. A systematic exploration, with various training and test data combinations, was conducted to isolate factors and assess the performance of ML relative to these factors. Multiple performance metrics (classification accuracy, true positive rates, false positive rates) were used to assess the ML model's accuracy in flaw classification.

Findings to date indicated that ML is capable of relatively high accuracy with respect to flaw classification. However, flaw type, size, and location were critical factors affecting the model classification accuracy on a test data set, with the accuracy varying depending on the flaw size and location. Given that small flaws are also challenging to detect and disposition accurately in manual analysis, the results may indicate inherent challenges with detecting small flaws in the vicinity of welds. The results also indicated the potential for learning key features of flaws and non-flaws using data from simple reflectors (saw cuts) while generalizing well to other flaw types as long as there are no confounding factors (such as inspection through welds). The results also pointed to the need for using multiple metrics to evaluate ML performance, including true positive rates, false positive rates, and false negative rates, to ensure that characteristics such as bias in the results can be identified. Preprocessing procedures should be consistent across the data used for training and testing ML, with inconsistencies in these stages having a major impact on classification accuracy.

The results pointed to the need for common, representative data sets for evaluating the performance of ML, with similar data sets likely needed for future qualification of these techniques. The ability to retrain and tune ML models to improve performance points to the need for robust procedures for verifying and validating ML models, including after any retraining, to ensure confidence in the results. Note that the analysis to date did not explicitly quantify the uncertainty in the ML classification results, and approaches

that estimate confidence in the classification result will likely be helpful in interpreting the classification outputs from ML models.

Ongoing research in this area continues the evaluation of ML with additional data, including the use of data augmentation techniques and the incorporation of simulation data with empirically derived data for training the algorithms. Applications beyond simple classification (including image segmentation) are also being assessed, and the findings are being used to develop recommendations for validation and qualification of ML prior to field use.

1. INTRODUCTION

Periodic inservice inspection (ISI) using nondestructive examination (NDE) 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, the rules and acceptance criteria for ISI are specified in Section XI (Rules for Inservice Inspection of Nuclear Power Plant Components) of the American Society of Mechanical Engineers (ASME) Boiler and Pressure Vessel Code (BPVC) ("Code"). In addition, some inspections are conducted to meet industry program (Materials Reliability Program [MRP] and Boiling Water Reactor Vessel Internals Program [BWRVIP]) requirements. 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 components.

Ultrasonic testing (UT) is one of the most common methods for inspecting the structural and non-structural components in nuclear power systems. For example, ultrasonic NDE has been widely used for the inspection of weldments in pipes and vessels. In addition to ultrasonic data collection using qualified equipment, the inspection process includes data interpretation for flaw detection, classification, and localization. An experienced and qualified NDE technician is needed for both equipment operation and data analysis. However, there is an expected shortage of qualified NDE technicians, especially in the nuclear power industry in the United States. Additionally, manual analysis of the data for flaw characterization is time-consuming, and the results can be subjective based on the technician's experience and knowledge.

As computer processing power continues to grow while the pool of qualified NDE inspectors dwindles, there has been a growing desire within the nuclear power industry to automate the processing of NDE data. This shift is made feasible by the strides in computational capabilities, the ready access to cloud-based computing, and the emergence of machine learning (ML) algorithms. Much like in other fields of science and engineering, ML, especially deep learning (DL), is seen as a promising avenue for automating NDE data analysis and performing routine and repetitive tasks. The growing number of recent publications on ML applied to NDE reflect the wide spectrum of approaches that are applicable to the problem. Notably, the industry has been making headway in employing ML algorithms to train computer systems in the identification of anomalies within NDE datasets. For example, in steam generator tube inspections, automated data analysis is now used to support human analysts [3], leveraging the wealth of data gathered from steam generator inspections.

In the context of the nuclear industry, ML algorithms are likely to be deployed in one of several ways:

1. Oversight assistance (assisting site personnel in reviewing results).
2. Analyst assistance (identify regions of interest that are then subject to human analyst review and dispositioning)
3. Fully automated (identify Region of Interest (ROI) and classify as flaw/non-flaw. In this case, secondary analysis by a human analyst may be possible)
4. Fully automated (analyst review only if flaws are identified)
5. Fully automated (site oversight review only of flaws identified)

Options 1 and 2 are likely the most near-term scenarios for the deployment of ML in NDE, based on publicly available information. While the use of ML for NDE does not directly impact plant automation and control, options 1 and 2 generally fall within the boundaries of autonomy levels 0 and 1 that are described in NUREG-2261 [4]. The other deployment options appear to align reasonably well with the higher levels of autonomy described in NUREG-2261. While deployment through options 1 and 2 will

require appropriate qualification methods to ensure confidence in the ML algorithm, it is likely that the bar for evidence of ML accuracy and performance will be higher for options 3-5.

The rapid growth in ML methods and the diversity of possible approaches indicate a need to assess the current capabilities of ML supported NDE data analysis and to identify any gaps or shortcomings in current ML technologies. Research to address identified gaps is likely to be valuable in developing the technical bases for guidance on the application of ML and for understanding the impacts of proposed Code activities related to the applicability of ML for NDE of nuclear power plant (NPP) components.

This report documents the results to date of the assessment of ML for ultrasonic NDE. The initial phase included a literature review, the results of which were disseminated through a peer-reviewed journal article [5]. Section 2 provides an introduction to the background information related to ML terminology. Section 3 of this report summarizes the methodology used for the literature review, along with the findings. In parallel, an empirical assessment of ML was conducted using available ultrasonic data sets. Section 4 describes the datasets used for the empirical assessment and the resulting reference or common data set that is currently being compiled. This comprehensive dataset includes data from a range configuration (probe, refracted angle, near-/far-side, frequency) on different types of flaws [for instance, saw cuts, thermal fatigue cracks (TFC), electrical discharge machined (EDM) notches] from multiple specimens, including stainless-steel austenitic welds and dissimilar metal welds. Section 5 of this report describes the ML models used in our research and the performance metrics selected for evaluating these models. In Section 6, we present the results from different combinations of data used for model training and testing and discuss insights from these results in Section 7. Finally, Section 8 summarizes the document with recommendations and a brief overview of ongoing and planned research.

2. BACKGROUND: MACHINE LEARNING

This section provides an introduction to commonly used ML terminology.

Artificial intelligence (AI) is a branch of computer science that focuses on creating systems and machines capable of performing tasks that typically require human cognition [6]. These tasks can include learning, problem-solving, decision-making, language understanding, and perception. **Machine learning** is a subset of AI that involves the development of algorithms and statistical models that enable computers to learn from and make predictions or decisions based on data. Instead of being explicitly programmed, machine learning systems use patterns and insights from data to improve their performance on specific tasks. **Deep learning** (DL) is a subset of ML in which multi-layered neural networks learn from large data sets.

While a variety of applications of ML are possible, most common applications may be categorized into problems associated with classification, segmentation, or regression. **Classification** refers to the use of ML to classify signals as belonging to one of multiple classes. In the specific case of NDE, the signals may be A-scans (time-series data), B-scan images, etc., with class labels such as no flaw, cracks, geometry, etc.

Segmentation, usually in the context of image analysis, refers to the use of ML or statistical methods to label individual image pixels into one of (usually) two classes (for instance, "non-flaw" and "flaw"). This approach differs from the classification problem in the potential number of classes and the type of input data that is typically used. Note, however, there are no fixed boundaries between segmentation and classification, and the approaches and descriptions may be used interchangeably.

Regression refers to the use of ML to learn relationships between input and outputs with the goal of predicting continuous valued output variables from a given set of inputs. In NDE, regression-based models may be used, for instance, in estimating flaw depth or length from a set of NDE measurements.

Anomaly detection is a particular approach to data analysis that is typically applicable to screening data. In this scenario, the algorithm is usually trained to predict data from a single class. Data that are outside this class (so-called out-of-distribution) are identified as anomalous. A number of variations on this approach are possible, depending on the specific algorithm used. Similarities exist with image segmentation, with the differences largely in the types of algorithms used. As between segmentation and classification, there are no hard and fast boundaries between segmentation and anomaly detection.

Model training in machine learning is the process of teaching a machine learning algorithm or model to make predictions or decisions by exposing it to a dataset (**training dataset**) containing examples. In **supervised learning**, the training dataset includes input data (either the raw or unprocessed measurements, or features computed from the measurements) and corresponding output labels or target values. **Unsupervised learning**, on the other hand, uses input data with no label information, and the training process is used to cluster the data. A third learning paradigm, **reinforcement learning**, is concerned with learning optimal actions to maximize a cumulative reward (mathematical function that computes the value or utility of the result of a proposed action), and is often used for optimization (for instance, optimal design) and data-driven control. Reinforcement learning also uses unlabeled data for this purpose and typically requires a model or environment of the problem to compute the rewards associated with actions.

In all learning paradigms, the training process involves adjusting the model's parameters or **weights** iteratively to minimize the difference between its predictions and the actual target values in the training data. Before training can be executed, an ML model is built and the model structure (**architecture**) and **hyperparameters** are identified. Hyperparameters in ML are parameters that are not learned from the training data but are set prior to the training process. An example of hyperparameters is the learning rate that helps adjust how quickly the model learns from data. Selection of too low a rate results in slow convergence of the model to the optimal weights that minimize the difference between its predictions and the actual target values in the training data. However, too high a learning rate will result in the model oscillating around the optimal set of weights. As such, hyperparameters such as the learning rate play a major role in controlling the learning process. The model structure and hyperparameters are generally adjusted to achieve the optimal performance of the model, and this process is called (**model** or **hyperparameter) tuning**.

During the training process, the **validation dataset** serves as an independent dataset to assess the model's performance. The model's performance on the validation dataset helps in tuning hyperparameters, avoiding **overfitting**, and determining if the model is generalizing well to new, unseen data. Overfitting is a phenomenon where the model learns the training data too well, to the point that it begins to capture noise or random fluctuations in the data. As a result, the overfitted model performs exceptionally well on the training data but fails to generalize effectively to new, unseen data.

Model training requires the specification of the loss function which defines the error between the model prediction (model output) and the desired output value. Common loss functions used in machine learning applications include the mean square error (MSE), mean absolute error (MAE), and binary cross-entropy loss, with the choice dependent on the problem type (classification vs regression) and characteristics. Each loss function penalizes the model to varying degrees if the model prediction deviates from the desired output value. Model training typically involves multiple **epochs**, with the model's parameters gradually adjusting to minimize the training error and improve performance. An epoch is a single pass through the entire training dataset during the training of the ML model. One epoch is complete when the

model has seen and learned from all the training data. Depending on the specific algorithm used for learning (for instance, gradient descent or stochastic gradient descent), the adjustment of the model parameters might occur after the ML model has seen all the training data (batch), some of the training data (mini-batch), or one example from the training data. Various algorithms (such as the RMSProp optimizer or ADAM optimizer) may be used to adaptively change the learning rate to improve the convergence speed of the model weights to the optimal values, with the choice of optimizer algorithm dependent on the computational constraints of the user. Once the model is trained, the model will be tested on a **testing dataset**. **Model testing** or **model inference** refers to the process of evaluating the **generalization performance** to see if the model can make accurate predictions or decisions on new, unseen data that it hasn't encountered during the training process.

A wide variety of ML models, supervised or unsupervised, exist and have been proposed and applied to classification and regression problems. A comprehensive survey of algorithms is available in the literature, and variations of these algorithms are being proposed. Many of these ML algorithms have been applied for NDE problems (see Section 3 and references there), and many more have the potential for application to NDE. The diversity in ML methods and their applications for ultrasonic data analysis raises questions about the applicability of ML-based data analysis for ultrasonic NDE as well as the contribution of ML analysis to improving the reliability of ultrasonic NDE. The assessment of NDE reliability when using ML algorithms, along with the various factors influencing ML analysis performance when applied to ultrasonic NDE, does not appear to have been examined in previous research.

A prototypic ML model needs to be selected since it is impossible to evaluate all proposed ML algorithms. In this work, the **Convolutional Neural Network** (CNN) was used as the prototypic machine learning model. A CNN is a specialized type of artificial neural network designed for processing and analyzing visual data, such as images and videos. Convolutional neural networks use a feedforward structure, where inputs to the network are transformed by successive layers of neurons. While many layer choices are possible, a typical CNN includes convolutional layers, max-pooling layers, and flattening layers. A fully connected layer is usually included at the end and is dedicated to the classification task. The max-pooling layer operates on each feature map separately and performs downsampling by dividing the input into rectangular pooling regions and taking the maximum value from each of these regions. Flatten layers reshape the output of the preceding convolutional or pooling layers into a one-dimensional vector, essentially collapsing or "flattening" the multi-dimensional (tensor) output of these preceeding layers into a single long vector. A fully connected layer is a traditional neural network layer where each neuron is connected to every neuron in the preceding and succeeding layers. Dropout and batch normalization layers are typically included and enable the CNN to learn without overfitting. Dropout randomly removes a subset of nodes temporarily from the network during each training epoch, forcing the network to learn the data using a smaller network. Batch normalization layers apply a transformation to the data to maintain the mean output of the network close to zero and the output standard deviation close to 1. These layers force the neural network to be robust to small variations or noise in the data, and lower the chances of the network overfitting to a small training data set. Activation functions (e.g., rectified linear units [ReLU], or sigmoid) are used in each neuron or unit in a layer, and are mathematical functions applied to the output of each neuron in the convolutional neural network prior to passing the output to the next layer in the network. While other activation functions can be used, the ReLU and sigmoid functions are widely used in image analysis applications [23].

CNNs are particularly well-suited for tasks like image classification, object detection, facial recognition, and image segmentation. Their applicability for such tasks stems from the structure of the CNN, which includes convolutional layers and pooling layers in a feed-forward (i.e., from input to output only, with no backward connections) architecture. Collectively, these types of layers allow for the use of raw measurements as inputs and the use of local information, such as flaw responses or scattering from grain boundaries in ultrasonic NDE data, to identify key features within the data. The use of convolutional

layers also provides some level of invariance to small shifts in the measurements, such as small shifts in the measured responses due to the position and tilt of the flaw within the specimen. Finally, the feedforward nature of the architecture allows for parallelization for training and inference purposes. Convolutional networks have shown success with classification tasks for ultrasonic NDT applications in recent years. Deep CNNs are able to learn from raw ultrasonic NDT signals (A-scan raw signals or image-like B-scan raw data) without the need for explicit feature engineering. The recent advances in machine models developed for other industries facilitate the application of ML for different NDT fields [7].

This work focused on the application of ML (e.g., CNN model) for flaw classification. **Flaw classification** refers to the use of ML to classify signals (A-scans, B-scan images, etc.) as belonging to one of multiple classes (non-flaw, cracks, geometry, etc.).

The rest of this report presents the key takeaways from the literature review and the results of an empirical assessment of ML data analysis for the classification of ultrasonic NDE, focusing on the factors influencing ML classification performance. While the research results presented use CNN, the assessments are expected to be applicable to other ML methods as well. The results of these assessments are expected to be useful to the U.S. NRC in developing guidance on the use of ML for NDE, as well as in future ASME Code activities.

3. LITERATURE REVIEW SUMMARY

Machine learning algorithms have been used for automated detection, classification, and localization of welding flaws. These ML methods include the kernel method, ensemble method, neural network method, and unsupervised method. This section briefly discusses the literature review on ML for ultrasonic NDE, including the review methodology used and related findings. Details can be found in the published journal article [5].

3.1 REVIEW METHODOLOGY

The literature about ML applications for ultrasonic NDE of welds was reviewed. For this effort, it was assumed that ML includes an element of learning from data unlike automated analysis methods that use classical statistical techniques that do not learn from data. Although ML is also widely applied to structural health monitoring (SHM) [8–10] and conditional assessment (CS) [11,12], this literature review focused on the ML applications for ultrasonic NDE of welds. Background information on ultrasonic NDE and ML may be obtained from multiple sources [13–15]. 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. In 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 (saw cuts, notches, and cracks) and non-flaws. Weld-fabrication flaws (for instance, lack of fusion, slag inclusions, and porosity) and signals from geometrical features such as counterbores were not explicitly considered in this study. The NDE methods of interest to this assessment were limited to ultrasonic testing (UT) with longitudinal and shear wave modalities, 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.

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) database. Keywords included ultrasonic nondestructive testing and evaluation (NDT&E), weld

defect/flaw, NDE automated analysis, ML, and DL for NDE. Out of approximately 200 articles identified through this search, about 135 were determined to be distinct and appropriate for this effort. 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 for 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.

3.2 FINDINGS FROM THE LITERATURE REVIEW

The literature review indicated a generally applicable workflow for "intelligent" automated data analysis (Figure 3-1). Broadly speaking, 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 on the training data set (Section 2). 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 Figure 3-1 are applied optionally (dashed boxes), whereas others may use the same stage multiple times to meet different objectives. Modern DL algorithms often skip the feature extraction stage, preferring to let the algorithm extract an implicit representation of relevant features during the learning phase.

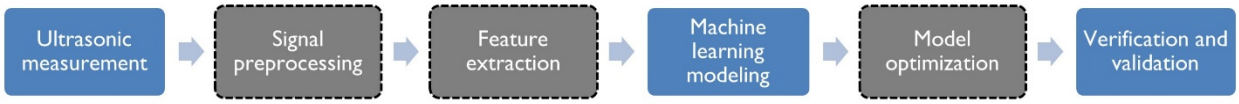


Figure 3-1. Summary of workflow for applying ML for UT signal classifications.

The key findings from the literature review are summarized below:

1) Feature Extraction and Selection

The literature indicates that ML can be successfully applied to A-scans, B-scans, and D-scans, and various feature extraction methods can be used to meet the analysis objectives. The applicability of several different types of feature extraction methods is based on the type of ultrasonic measurement. 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. When using feature extraction as a preprocessing step, it is important to note that while 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 classification performance. Therefore, the selection of appropriate features for input is critical for building an efficient learning model with a high model performance.

2) Data Selection

The literature suggests that ML performance is mainly reliant on the size and quality of the data sets used in the ML model training. The question of sufficient sample size and quality, including the distribution of data across multiple classes, was not explicitly reported in the literature on ML for ultrasonic NDE. 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, porosity, non-flaw, etc.) is approximately the same. A slight imbalance in the training data across classes is generally not an issue for achieving well-trained ML models. However, severe data imbalance can result in less than desired ML model performance. Data bias is a type of error in ML in which certain classes are overweighted, and others are less weighted. Based on the literature surveyed, the extent of the impact of data bias on the reported classification performance for ultrasonic NDE is unclear.

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 process (e.g., optimization algorithm, loss function, learning rate, etc.). The results reported in a few articles that employed hyperparameter tuning indicate the value of such optimization for improving model performance. Hyperparameter tuning should be considered a best practice for the use of ML for ultrasonic NDE problems.

4) ML Model Verification and Validation

ML model verification and validation (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. The literature review results suggest that a separate data set not used in 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. A second issue identified earlier is the validation of the data itself to ensure that the data used to test the ML models and the data used to train them are in the same distribution.

The literature review indicated that ML methods may be applied to most inspection setups and that there are no inspection-related constraints on using 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 NPP inservice 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. While DL algorithms may be expected to achieve the best classification performance in ultrasonic NDE based on results in other application domains, the best-reported results do not appear to be from DL algorithms, and it is not clear whether this is an inherent issue with DL 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 ensuring 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 [16]. The 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 that the 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 in 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 review also identified that there was limited information in the literature on the application of ML to ultrasonic NDE 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, perhaps as part of essential variables associated with ML algorithms. 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 applied primarily to the standardized image and time-series data sets will eventually find their way into the NDE application domain.

4. ULTRASONIC NDE REFERENCE DATASET

Empirical assessment of ML defect classification performance requires a data set. While such analyses may be performed using any data, the compilation of a reference data set for such analysis enables a robust assessment of algorithms as well as various factors that may influence the results. The availability of a reference data set also allows an unbiased comparison of performance from different algorithms.

In this section, the reference dataset used in this work is briefly discussed, and the details of data compilation and preprocessing before data are fed into the ML models are provided. This data set is part of a larger common dataset built for future studies. The reference dataset was compiled from experimental ultrasonic inspections on multiple specimens and various inspection setups to assess the capability and limitations of ML models for ultrasonic NDE of welding defects. The data set includes data from the two types of artificial flaws—machined saw cuts and fabricated thermal fatigue cracks (TFCs)—that were introduced into the specimens, with inspection data from these flaws used to train and test the ML algorithms.

4.1 REFERENCE DATA: SPECIMENS

The dataset used in this work for the ML studies was part of a reference dataset of ultrasonic B-scan images. This dataset included the eight specimens in Table 4-1, consisting of six stainless-steel (SS) specimens and two dissimilar metal weld (DMW) specimens. The dataset was collected using multiple types of ultrasonic probes, including conventional longitudinal and shear probes [GEIT 2MHz (transmit-receive longitudinal transducer, GE Inspection Technologies, Niskayuna, NY), SNI 2MHz (transmit-receive longitudinal transducer, Sensor Networks, Inc., Boalsburg, PA), and SwRI 2.25MHz (single-element shear transducer, Southwest Research Institute, San Antonio, TX)] and the phased array probe (PAUT 2MHz). Different refracted angles were used with these transducers: 45˚, 60˚, and 70˚. The data were also acquired with a skew angle of 0˚ and a skew angle of 180˚. Table 4-2 summarizes the different setup combinations of each specimen and probe. The six stainless-steel specimens were inspected using all four probes, all three refracted angles, and two skew angles. The two DMW specimens only have the data from a skew angle of 180˚ for different probes and refracted angles. Additional details about the specimens and flaws are included in Appendix A. A subset of these data was used in the empirical assessment described in this document.

Table 4-1. Summary of the specimen and flaw information in the common dataset.

Specimen ID	Description	Base Material	Material Class	Weld Material	Flaws	Geometry
02-24-15	24 in. Sch80 304 SS	A358	SS	Not provided	TFC, SC	Pipe
322-14-01P	14 in. Pipe to CSS Valve	316	SS	316 SS	TFC	Pipe
		SA351	CSS	Not provided		
19C-358-1	Custom SS Plate	304	SS	308 SS	SC	Plate
19C-358-2	Custom SS Plate	304	SS	308 SS	SC	Plate
21C-303-1	Custom SS Plate	304	SS	308 SS	EDM notch, TFC	Plate
21C-303-3	Custom SS Plate	304	SS	308 SS	EDM notch, TFC	Plate
8C-032	DMW	A321	CS	309 SS	TFC	Pipe
		316	SS	Inconel 182		Safe End
		SA508	CS	/		Nozzle

8C-091	DMW PZR Surge Nozzle Specimen	/	CSS	/	EDM notch, TFC	Pipe
		/	CS	/		Nozzle

Table 4-2. Summary of specimen and inspection setup combinations in the common dataset

Probe	GEIT 2MHz				PAUT 2MHz						SNI 2MHz				SwRI 2.25MHz					
Refracted angle	45˚		60 ˚		45 ˚		60 ˚		70 ˚		45 ˚		60 ˚		45 ˚		60 ˚		70 ˚	
Skew angle	0	180	0	180	0	180	0	180	0	180	0	180	0	180	0	180	0	180	0	180
02-24-15	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
322-14-01P	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
19C-358-1	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
19C-358-2	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
21C-303-1	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
21C-303-3	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
8C-032		x		x		x		x		x		x		x		x		x		x
8C-091		x		x		x		x		x		x		x		x		x		x

4.2 EMPIRICAL ASSESSMENT: SPECIMENS AND ULTRASONIC INSPECTION SETUP

1) Specimens and Flaws

For the ultrasonic inspection data collection process, four stainless steel specimens and two DMW specimens were used. All the stainless-steel specimens had stainless steel bases and weld materials. Among them, 19C-358-1 and 19C-358-2 were plates, 322-14-01P was a pipe section, and 02-24-15 was a whole pipe specimen. DMW specimens were represented by specimens 8C-032 and 8C-091. Detailed information on these six specimens can be found in Table 4-3.

Specimens 19C-358-1 and 19C-358-2 each contained four saw cuts (SCs), while 322-14-01P had three TFCs. Specimen 02-24-15 had three TFCs and four SCs. 8C-032 had four TFC, while 8C-091 had two electrical discharge machined (EDM) notches and two TFC flaws. Overall, there were 12 SC flaws, 12 TFC flaws, and two EDM notch flaws among these six specimens. The flaw lengths and heights, expressed as a percentage of the specimen thickness, are also detailed in Table 4-3. The flaw length and relative height information were plotted in Figure 4-1. The length of the flaws varied from 10.7 mm to 101.7 mm, while their height ranged from 7.5% to 65.8% of the specimen thickness.

2) Ultrasonic Inspection Setup

Ultrasonic data were collected using a Zetec DYNARAY system with longitudinal transducers, shear mode transducers, and phased array probes. For this particular investigation, the data obtained from the 2.25 MHz SwRI shear transducers were used), along with some data from the GEIT longitudinal transducer. Different refracted angles with the shear and longitudinal mode transducers, such as 45˚, 60˚, and 70˚, were employed along with the transducers.

Figure 4-2 displays the top view positions of the four SC notches on 19C-358-1 as an example of the specimens included in the reference data set. The weld center line is marked in the figure as a green dash line. Thus, Notch 1 is located away from the weld center line in the HAZ, while Notch 4 is situated close

to the weld center line. Notch 2 and Notch 3 were partly in the weldment. In the figure, the scanning was conducted in the X- and Y- directions on the specimen's top surface.

Table 4-3. Summary of the flaw information

Specimen	Description	OD to ID Thickness (mm)	Flaw	Type	Flaw length (mm)	Height (% thickness
19C-358-1	Custom SS Plate	73.1	1	Saw cut	101.7	30.1%
			2	Saw cut	101.4	30.2%
			3	Saw cut	101.6	30.2%
			4	Saw cut	101.4	30.0%
19C-358-2	Custom SS Plate	28.6	1	Saw cut	100.6	29.2%
			2	Saw cut	101.4	29.2%
			3	Saw cut	101.4	29.4%
			4	Saw cut	101.4	29.5%
322-14-01P	14 in. Pipe to CSS Valve	38.6	1	TFC	70.4	65.8%
			2	TFC	13.5	12.5%
			3	TFC	46.5	43.0%
02-24-15	24 in. Sch80 304 SS	36.0	A	TFC	10.7	15.0%
		35.3	B	TFC	30.5	43.0%
		35.7	C	TFC	43.6	64.0%
		36.0	a	Saw cut	32.8	7.5%
		35.9	b	Saw cut	65.2	28.4%
		36.2	d	Saw cut	54.1	18.8%
		35.8	e	Saw cut	43.7	12.0%
8C-032	DMW pipe	38.2	1	TFC	22.9	20.0%
		36.0	2	TFC	28.9	40.0%
		38.2	3	TFC	45.9	60.0%
		36.0	4	TFC	21.6	30.0%
8C-091	DMW PZR Surge Nozzle Specimen	38.6	1	EDM notch	69.1	30.2%
		39.9	2	EDM notch	50.8	17.6%
		38.8	3	TFC	70.6	36.4%
		39.7	4	TFC	57.6	23.2%

Scans on the -Y side (right side of the weld center line) were designated as skew angle 0˚, with scans on the +Y side of the centerline designated as skew angle 180˚. In this case, the 0˚ skew angle inspection indicates that the probe is on the same side of the weld as Notches 1, 2, and 3, and the data could be considered as near-side data. The data from a skew angle of 180˚ may be considered as far-side examination data for these flaws.

Figure 4-3 shows an example of the B-scan image, which was collected at the middle of Notch 1 (see black dash line in Figure 4-2) on 19C-358-1 with a 45˚ refracted angle and 0˚ and 180 ˚ skew angle. In Figure 4-3 (a) (near side B-scan image), the horizontal direction is the Y-direction in Figure 4-2 and the Z-direction is the thickness direction of the specimen. The figure shows a strong reflection from the bottom of the SC and a weak feature (tip diffraction) from the top of the SC. The right-side boundary of the weldment can also be seen in the B-scan image. If the B-scan image was collected from the far-side, the B-scan image [Figure 4-3 (b)] would contain more noise (lower signal-to-noise ratio) than the image collected from the near side.

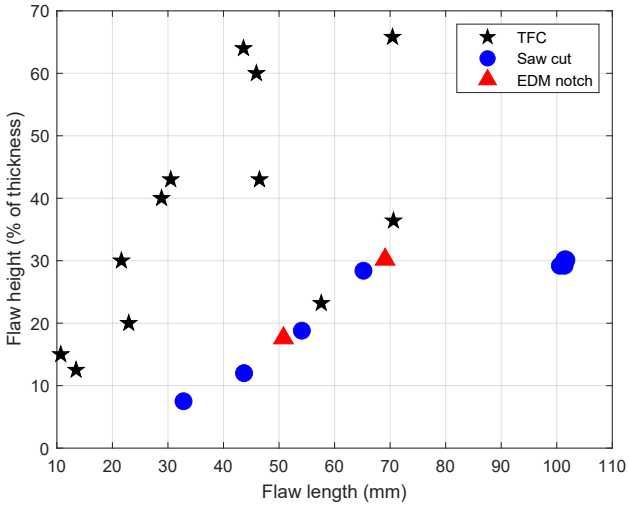


Figure 4-1. Flaw size distribution for the four stainless steel and two DMW specimens.

Different hardware gains may be used during the inspection. The gain used for each test is listed in the last column of Table A2 in Appendix A and was used for gain adjustment in the image preprocessing (described in Section 3.3).

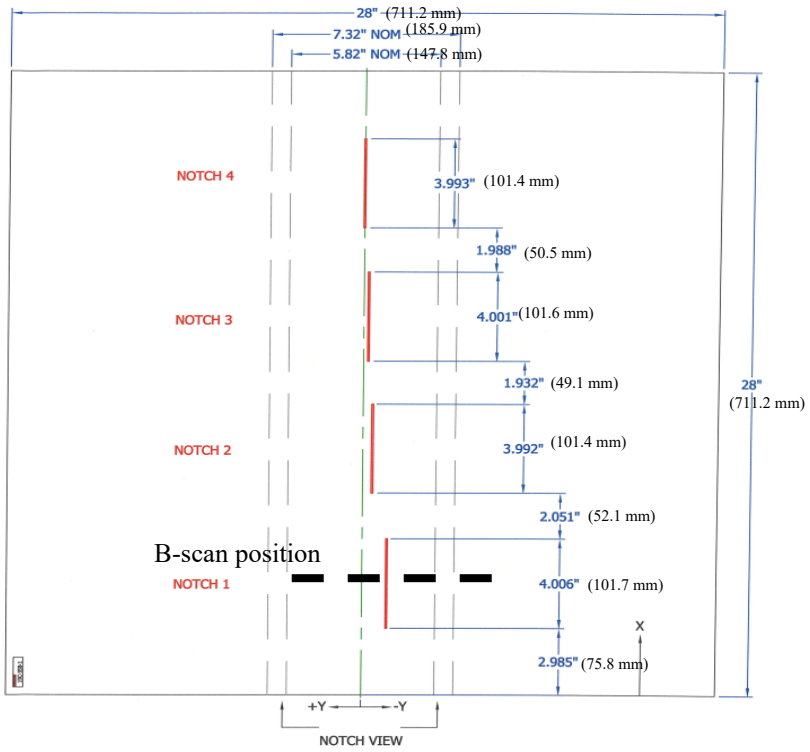


Figure 4-2. Top view of specimen 19C-358-1 with four SC notches.

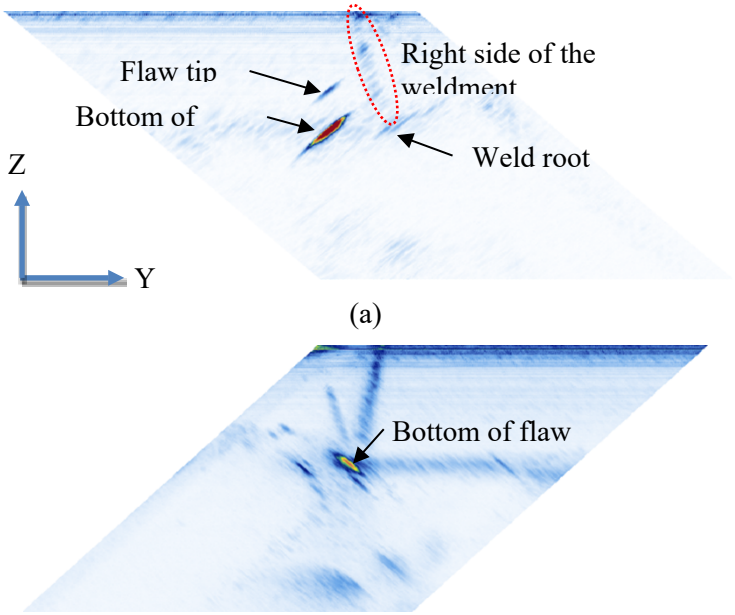


Figure 4-3. B-scan image of flaw 1 on specimen 19C-358-1 using SwRI 45° probe: (a) near side, (b) far-side

4.3 PREPROCESSING OF THE B-SCAN IMAGES

The inspection data were stored as a three-dimensional matrix, with the X-direction corresponding to scans parallel to the weld, the Y-direction perpendicular to the weld, and the Z-direction representing the depth direction. B-scan images represent cross-sectional slices in the Y-Z plane. Before input to the ML models, the B-scan images were preprocessed with gain adjustment, image cropping, and down-sampling.

1) Gain Adjustment and Amplitude Scaling

Preliminary studies indicated that the ML classification accuracy was impacted by the use of data collected with differing hardware gain values. Since different hardware gains were used during the data collection on the different flaws, a gain adjustment step was performed for the B-scan images to normalize all the data to the same range. Specifically, the B-scan data was scaled, so it corresponded to an acquisition gain value of 40 dB (the approximate median value of all the hardware gain values in the data set). Further, the B-scan data were rectified and scaled to a range of 0 to 1 after the gain adjustment, as part of the data normalization process prior to applying as inputs to the ML model. Data normalization enables the ML model to learn faster and is a recommended step of any input preprocessing work flow. Figure 4-4(a) shows a B-scan acquired with original gain of 28 dB, while Figure 4-4(b) shows the B-scan after gain adjustment to 40 dB.

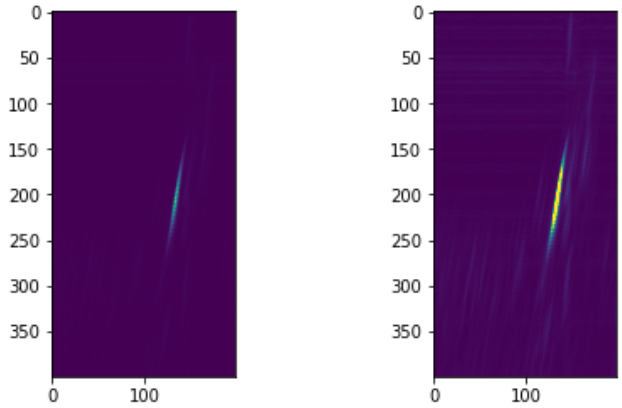


Figure 4-4. B-scan images with different gains: (a) gain of 28 dB, (b) gain of 40 dB after gain adjustment.

2) Image Crop and Down-Sampling

The size of each B-scan image in the data set ranged from 4300 to 6700 data points in the Z-direction and from 217 to 401 points in the Y-direction. These differences arise from differences in the scan direction (Y-direction) and the specimen thickness (Z-direction). The size of the B-scan images was too large to handle in the ML models due to limited GPU memory. Further, the B-scan image contains surface noise and flaw signals corresponding to a full skip or single Vee reflection [Figure 4-5(a)]. Such noise was seen to impact the ML results in preliminary studies, likely due to the introduction of unrelated information. Therefore, the front surface noise was removed from the image first, with the top 500 to 1000 data points in the Z-direction cropped from the B-scan image in Figure 4-5(b). Additionally, 1000 to 2000 data points in the bottom portion of the image, which contains multiple skip reflections and noise, were also cropped. Finally, to reduce the computational cost, the cropped B-scan image was downsampled to a standard size [400 (Z)×200 (Y)] for all B-scan images. Note that other approaches for B-scan image size reduction are also proposed in the literature, such as max pooling with ¼ wavelength (λ) to reduce the B-scan image size without losing information from the sound path [17].

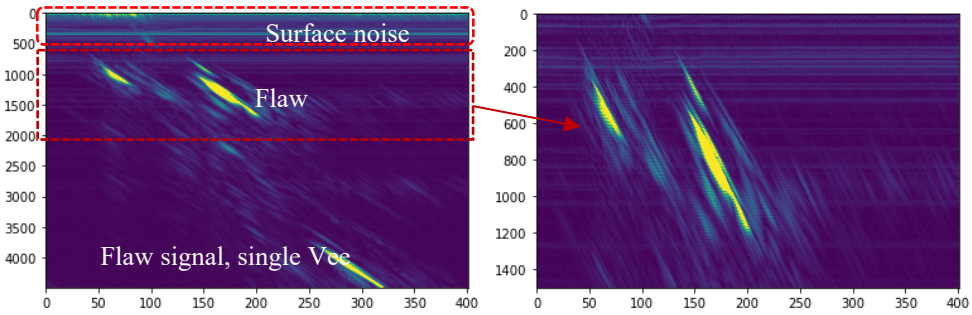


Figure 4-5. (a) Original B-scan image with surface noise and bouncing reflection, (b) B-scan image after cropping.

3) Orientation Correction: Image Reflection

For most specimens, the data from skew angle 0˚ is the near-side scanning data. However, for certain specimens (such as 19C-358-2), data from a skew angle of 180˚ corresponded to the near side. Figure

4-6(a) and (b) show the B-scan images of flaw 1 on specimens 19C-358-1 and 19C-358-2, respectively. The flaw features in these two images demonstrate different orientations. Variations in the characteristics of the data used to train ML algorithms can make training the algorithms challenging. To ensure a common orientation within the reference data set, images such as those in Figure 4-6(b) were reflected in the X-direction. An example of the orientation-corrected data is shown in Figure 4-6(c).

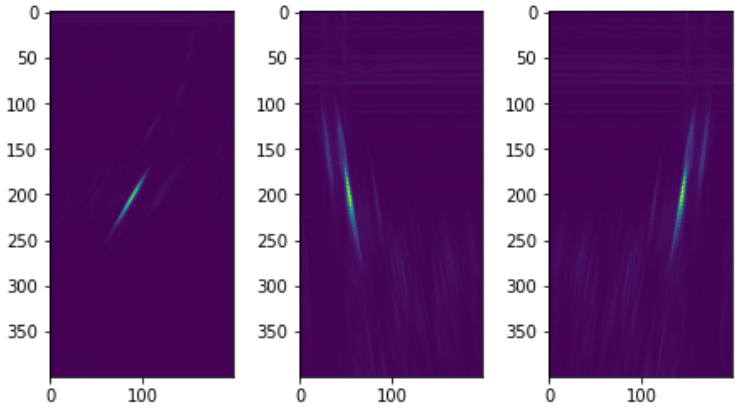


Figure 4-6. B-scan images of (a) flaw 1 on specimen 19C-358-1 with skew angle 0°, (b) flaw 1 on specimen 19C-358-2 with skew angle 180°, (c) flipped image of Flaw 1 on specimen 19C-358-2 with skew angle 180°.

4.3.1 Label Assignment

As discussed in Section 2, the use of supervised ML methods requires labels be assigned to each input. In this study, the labels corresponded to two classes – Flaws and Non-Flaws. Manual review of the data indicated that slight shifts in the inspection scan's origin resulted in potential inaccuracies in labels assigned automatically based on flaw positions. The potential label assignment inaccuracies were associated with the B-scans at the edges of the flaws (Figure 4-7). As part of the manual assignment of labels the decision was made to assign a "transition" class label to B-scans at the edge of each flaw (Figure 4-8). Such a designation simplifies the selection of data for training, since data from the flaw edge can be removed from the training data set, or assigned to the Flaw class and used to train so-called "conservative" ML models that are likely to flag more indications as flaws. Alternatively, a portion of the flaw edge data (say corresponding to below a 6 dB flaw amplitude threshold) can all be assigned to the Non-Flaw category to assess the impact of similar data in two classes on the ML classification accuracy. Results reported in this document chose the first approach (removing the flaw edge data from the training data set); future studies will focus on the other approaches. Preliminary analysis indicated that mislabeling of the data had the potential to degrade the classification accuracy by 20% or more on the data used in this study, highlighting the importance of accurate labeling in the training data set.

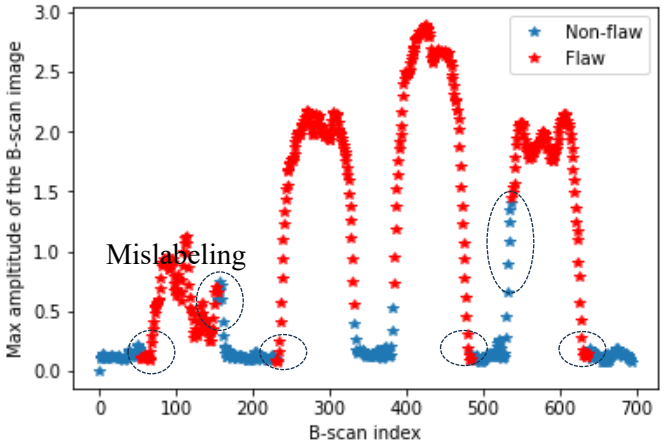


Figure 4-7. Mislabeling of the B-scan images for specimen 19C-358-2.

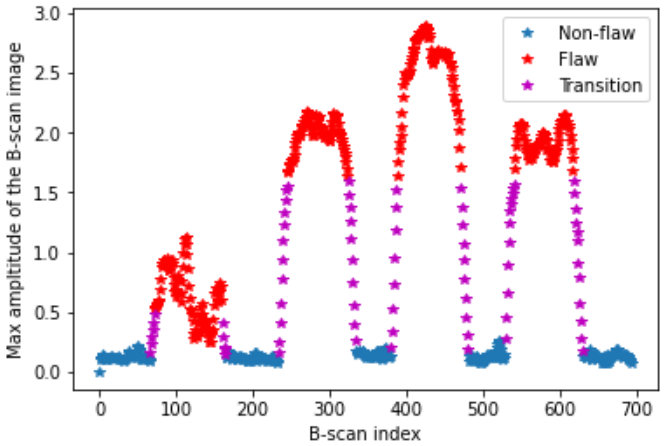


Figure 4-8. Labeling after correction of the B-scan images for specimen 19C-358-2.

5. MACHINE LEARNING MODELS

Most of the analysis conducted in this study used convolutional neural networks (CNN). CNNs and their variations are among the commonly used ML (specifically DL) models for image analysis (segmentation, classification). Given their ubiquity in image analysis applications, findings from research using CNNs as the prototypic model are expected to be broadly applicable to other models.

In this section, details of CNN models used in this study are discussed. To identify the factors influencing the ML model performance, several test matrices with various training and test data combinations were developed and are listed. Finally, metrics for ML performance evaluation are described.

5.1 CONVOLUTIONAL NEURAL NETWORK

A CNN is a specialized type of deep neural network (DNN) primarily designed for analyzing images. Unlike many other image classification algorithms, CNNs require minimal preprocessing of input data. CNNs have found extensive applications in defect classification, particularly in using ultrasonic NDE time domain signals and B-scan images [18–21]. In the context of a CNN (Figure 5-1), the convolutional layer and the pooling layer collaborate as feature extraction components, while fully connected layers are employed for the classification task. This architectural arrangement empowers CNNs to focus initially on low-level features, which are subsequently integrated to form higher-level features in subsequent layers. Figure 5-1 depicts an illustrative diagram of a CNN architecture using B-scan images as input for the purpose of flaw classification.

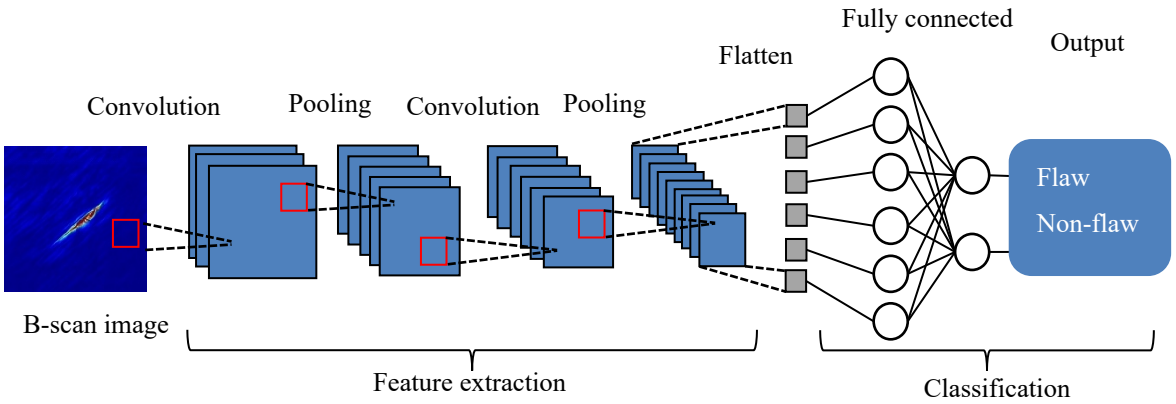


Figure 5-1. Example architecture of a CNN model.

The issue of overfitting poses a significant challenge during the training of deep machine learning models, particularly when the available training data are limited. As described in Section 2, 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. Previous research [12, 21] has demonstrated that introducing batch normalization layers or dropout layers can enhance the model's ability to generalize and mitigate overfitting concerns. However, it is crucial to exercise caution when incorporating both types of layers into the model architecture, as an unconstrained combination of the two layers may lead to reduced accuracy, as documented in a prior study [22]. In the course of this study, the architecture of a CNN was fine-tuned, resulting in the architecture depicted in Figure 5-2, building upon the framework originally proposed by Virkkunen et al. [7]. This CNN structure features four convolutional layers, three max-pooling layers, a flatten layer, and a fully connected layer dedicated to classification. The number of neurons (e.g., 96), convolutional filter size (e.g., 3×3), and activation function (e.g., ReLU) used are shown in the diagram for the convolutional, max pooling, and fully connected layers. Both dropout layers and batch normalization were integrated into the CNN's design, and extensive experimentation and tuning were performed using data from four specimens. The training used the binary cross-entropy loss function and a learning rate of 0.0001 along with the RMSProp optimizer. After conducting numerous rounds of tuning experiments, the optimal network performance was achieved by incorporating three dropout layers with a dropout rate of 0.5 within the network structure.

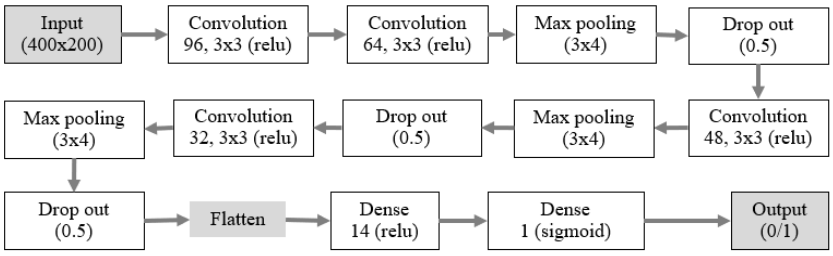


Figure 5-2. The architecture of the CNN model used in this work.

5.2 TRAINING AND TESTING MATRIX

The extracted B-scan images from the six specimens were labeled as flaw or non-flaw according to the B-scan position and the designed flaw position. For each setup type, such as SwRI 45˚ with a skew angle of 180˚, a comprehensive breakdown of the flaw and non-flaw B-scan images is summarized in Table 5-1. In the case of the aforementioned setup, the dataset comprised 2,373 non-flaw and 1,556 flaw B-scan images. Similar datasets with nearly equal numbers of flaw and non-flaw B-scan images were compiled for various setups, such as SwRI 60˚, SwRI 70˚, and GEIT 45˚. To facilitate the training and evaluation of the ML models, a portion of the data was designated for training purposes, while another subset was reserved for testing. The selection of data for training and testing hinged on the specific configuration within the training and testing matrix. Before initiating the training process, all B-scan images were preprocessed uniformly, as detailed in the preceding section.

Within the flaw images, three distinct types of flaws were identified: SC, TFC, and EDM notch, with the latter being considered a variant of the SC flaw. To ensure uniformity, all B-scan images were preprocessed and resized, resulting in consistent dimensions of 400×200 pixels. These processed images were subsequently used as input for model training and testing. During the training phase, 75% of the dataset designated for training purposes was dedicated to training the model, while the remaining 25% served as the validation dataset. The validation dataset plays a pivotal role in tuning the model hyperparameters and conducting interim assessments of model performance during the training process.

The co-mingling of test data from multiple specimens can lead to information leakage, where the ML model learns specific characteristics of the specimens instead of the flaws. To ensure robustness and prevent the model from learning features specific to a single specimen, testing data were sourced from specimens not used in the training dataset. Further, while the training data and validation data could originate from the same specimens, the training and validation data were sourced from different flaws, guarding against the model memorizing individual flaws rather than learning generalized features. It should be noted that the selection of specific flaws for inclusion in the training set was done randomly.

Different training and testing combinations are listed in Table 5-2 through Table 5-5 to study the factors that may affect classification performance. In Table 5-2, the objective was to examine the influence of data size and flaw type in austenitic weld specimens on classification performance. The ML model was trained using data from a single specimen, subsequently subjecting it to testing using data from all three other austenitic weld specimens, resulting in Test scenarios 1, 3, 4, and 6. Test 2 entailed training with data from two specimens (19C-358-1 and 19C-358-2), augmenting the dataset to assess the impact of data volume on model performance. Test 5, on the other hand, focused on the effect of specimen variability by exclusively using SC data from specimen 02-24-15 for training while keeping the flaw type and other variables constant. In Test 6, the performance of transfer learning was explored, wherein the model was

initially pretrained on TFC data, followed by fine-tuning with SC data, culminating in testing with SC data derived from the remaining two specimens.

Table 5-1. Summary of the flaw and non-flaw B-scan images for each setup (e.g., SwRI 45˚, skew angle 180˚)

Specimen	Flaw	Type	Number of B-scans Labeled as "Flaw"	Number of B-scans Labeled as "Non-Flaw"
19C-358-1	1-4	SC	406	288
19C-358-2	1-4	SC	406	288
322-14-01P	1-3	TFC	138	352
02-24-15	A, B, C	TFC	80	159
	a, b, d, e	SC	198	251
8C-032	1, 2, 3, 4	TFC	92	712
8C-091	1, 2	EDM notch	119	168
	3,4	TFC	127	155
Total			1,556	2,373

For the test sequences described in Table 5-3, the focus was on evaluating the impact of probe differences (longitudinal vs shear mode probes). The initial test involved training the model with data from specimen 19C-358-1, acquired using a SwRI 45˚ shear probe, and subsequently assessing its performance with data from multiple stainless-steel specimens collected using the longitudinal GEIT 45˚ probe. The second test mirrored this configuration but with the training data derived from the longitudinal probe and testing with shear probe data. Table 5-4 summarizes the tests conducted to study the influence of refracted angles on the classification performance. Three tests were designed systematically using data acquired at refracted angles of 45˚, 60˚, or 70˚ for model training. Subsequently, the model's testing phase uses data sourced from the other refracted angles. These types of exploration allow for an understanding of how generalizable the information in any single wave mode or inspection angle data may be and how broadly applicable the trained models would be. These analyses also provide insights into the need to incorporate data covering multiple wave modes and inspection angles as part of the training data set.

The tests listed in Table 5-5 were used to investigate the model's generalizability across different specimen types. Specifically, the ML model was trained on either austenitic weld specimens or DMW specimens and tested using data collected from the opposing specimen type. This analysis aimed to ascertain the model's capacity to generalize effectively across disparate specimen categories, providing insights on adaptability and transferability across these categories.

Table 5-2. Training and testing matrix for stainless steel specimens

Test	Training Specimen (flaw, type)	Testing Specimen (flaw, type)	Notes
1	19C-358-1 (1-4, SC)	19C-358-2 (1-4, SC)	Same flaw type
		322-14-01P (1-3, TFC)	Different flaw type
		02-24-15 (1-3, TFC)	Different flaw type
		02-24-15 (4-7, SC)	Same flaw type, but different sizes
2	19C-358-1 (1-4, SC) + 19C-358-2 (1-4, SC)	322-14-01P (1-3, TFC)	Double training data
		02-24-15 (1-3, TFC)	
		02-24-15 (4-7, SC)	
3	322-14-01P (1-3, TFC)	19C-358-1 (1-4, SC)	Different flaw type
		19C-358-2 (1-4, SC)	
		02-24-15 (1-3, TFCs)	Same flaw type
		02-24-15 (4-7, SC)	Different flaw type
4	02-24-15 (1-7, TFC and SC)	19C-358-1 (1-4, SC)	Training data has two types of flaws
		19C-358-2 (1-4, SC)	
		322-14-01P (1-3, TFCs)	
5	02-24-15 (4-7, SC)	19C-358-1 (1-4, SC)	Same flaw type, but different sizes
		19C-358-2 (1-4, SC)	
6	322-14-01P (1-3, TFC) retrain with 02-24-15 (4-7, SC)	19C-358-1(1-4, SC)	The model was first trained with C and retrained with D
		19C-358-2 (1-4, SC)	

Table 5-3. Training and testing matrix for different probes

Test	Training Specimen, Probe, Angle	Testing Specimen, Probe, Angle	Notes
1	19C-358-1 SwRI 45˚ (shear)	19C-358-1, GEIT 45˚	Same specimen, different probes
		19C-358-2, GEIT 45˚	Different specimens and probes
		322-14-01P, GEIT 45˚	Different specimens and probes
		02-24-15, GEIT 45˚	Different specimens and probes
2	19C-358-1 GEIT 45˚ (longitudinal)	19C-358-1, SwRI 45˚	Same specimen, different probes
		19C-358-2, SwRI 45˚	Different specimens and probes
		322-14-01P, SwRI 45˚	Different specimens and probes
		02-24-15, SwRI 45˚	Different specimens and probes

Table 5-4. Training and testing matrix for refracted angles

Test	Training Specimen, Probe, Angle	Training Specimen, Probe, Angle
1	19C-358-1, SwRI 45˚	19C-358-1, SwRI 60˚
		19C-358-1, SwRI 70˚
2	19C-358-1, SwRI 60˚	19C-358-1, SwRI 45˚
		19C-358-1, SwRI 70˚
3	19C-358-1, SwRI 70˚	19C-358-1, SwRI 45˚
		19C-358-1, SwRI 60˚

Table 5-5. Training and testing matrix for mixing stainless steel and DMW specimens

Test	Training Specimen (Flaws)	Testing Specimen (Flaws)
1	19C-358-1 (4 SC)	8C-032 (4 TFC)
		8C-091 (2 notches, 2 TFC)
2	02-24-15 (3 TFC & 4 SC)	8C-032 (4 TFC)
		8C-091 (2 notches, 2 TFC)
3	8C-032 (4 TFC)	19C-358-1 (4 SC)
		02-24-15 (3 TFC & 4 SC)
4	8C-091 (2 notches, 2 TFC)	19C-358-1 (4 SC)
		02-24-15 (3 TFC & 4 SC)

5.3 PERFORMANCE METRICS

The assessment of the model's classification performance evaluation used several pertinent metrics. These metrics include classification accuracy, confusion matrix, TPR, FPR, and the receiver operating characteristic (ROC) curve. Classification accuracy, a fundamental measure, quantifies the proportion of correct predictions relative to the total number of inputs, serving as a foundational indicator of model performance. The confusion matrix is a visualization tool (Figure 5-3), providing a tabular representation of the various outcomes stemming from the model's predictions in a classification problem. Within this matrix, four essential elements are identified - the number of: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

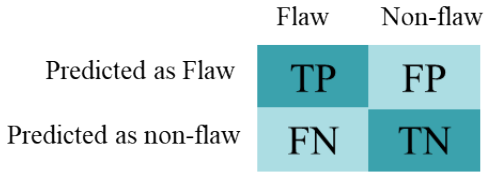


Figure 5-3. Example of a confusion matrix.

The classification accuracy may be defined using these four essential elements as:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}. \quad (1)$$

In addition to these measures, the true positive rate (TPR) and false positive rate (FPR) furnish valuable insights into the model's capabilities to correctly identify positive instances and the inadvertent misclassification of negative instances, respectively. Using the four elements in the confusion matrix, TPR and FPR could be calculated as below:

$$TPR = \frac{TP}{TP + FN}, \; FPR = \frac{FP}{FP + TN}, \quad (2)$$

where TP, FN, FP, and TN are the number of true positives, false negatives, false positives, and true negatives, respectively. In this work, the TPR represents the ratio of the number of correctly detected flaws to the total number of flaws. The FPR is the ratio of non-flaws that are classified as flaws to the total number of non-flaws. Note that the TPR and FPR may be expressed as a ratio (between 0 and 1) or as a percentage.

The ROC curve serves as a graphical representation of the model's performance, illustrating the tradeoff between the TPR and the FPR. This curve aids in the assessment of the model's discriminatory capacity and provides a visual summary of its classification effectiveness. When comparing ROC curves, better classification performance is shown as a higher TPR at the same FPR. The ideal condition, with a zero FPR and a perfect classifier, would correspond to the top left corner in the ROC curve plots.

The ideal condition for other metrics ((classification accuracy, TPR, FPR, etc.) would correspond to the perfect classifier (100% classification accuracy, with no false calls or missed calls). However, the perfect classifier is difficult to achieve in practice, and instead of the ideal condition, it is important to define acceptable values for the various metrics. Acceptable levels for the performance metrics (TPR, FPR, etc.) are likely to be problem-dependent. For instance, the desired or acceptable TPR for an inspection data screening task may be 100%, while that for a flaw classification problem may be lower. In all cases, the technical bases behind the acceptable levels for performance metrics will need to be established as these are likely to drive the performance requirements for qualification of ML algorithms for NDE data analysis.

6. PERFORMANCE ANALYSIS

Detailed results from the empirical evaluations of ML when using the reference data set are presented in this section. These results are categorized into three subsections by the data sources. Section 5.1 discusses the training and test data from the stainless-steel austenitic weld specimens. The main goal was to study the factors that influence ML performance at the flaw level (e.g., flaw type, location, dimensions, etc.). Section 5.2 describes the results, focusing on the influence of the inspection setup, including probe modes and refracted angles. In the last section (Section 5.3), the data from DMW specimens and stainless-steel specimens are mixed for training and testing to investigate the ML model's ability to generalize across specimens with different weld characteristics.

It should be noted that the results described here are from experiments that utilized a small set of empirically acquired measurements. Taken individually, these results might provide limited information but collectively, these results provide insights into the challenges/capabilities of ML for ultrasonic NDE. It is also worth noting that the results need confirmation with additional data, and studies to confirm and validate these results are ongoing.

6.1 TRAINING AND TESTING USING STAINLESS STEEL SPECIMENS

In the initial phase of the study, the CNN model was trained using data from specimen 19C-358-1. Within this training process, the subset of data corresponding to flaw 4 in the 19C-358-1 specimen was set aside as the validation dataset. After 200 epochs of training, the model exhibited a classification accuracy of 0.95 on the training dataset, while the accuracy on the validation dataset was determined to be 0.81. While not especially high, the classification performance on the validation data set is generally used to ensure that the ML model did not overfit on the training data. A better estimate of the ML accuracy may be obtained from its performance on the test data.

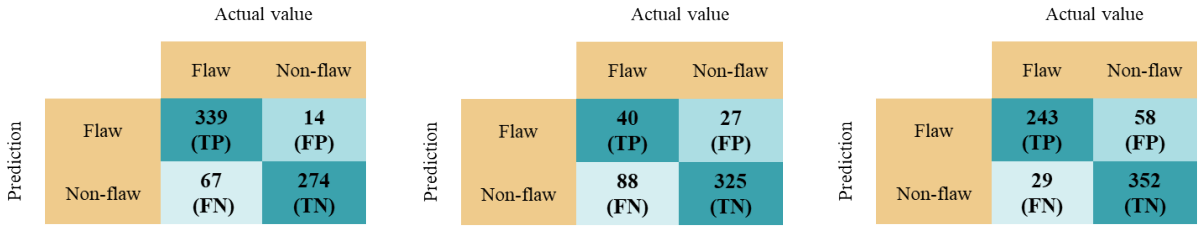
The model was tested with a dataset from the other three stainless steel specimens. The results of this evaluation were presented with the confusion matrices in Figure 6-1. Furthermore, the details of the breakout test results have been tabulated in Table 6-1.

6.1.1 Impact of Flaw Size and Location

The trained model showed a high classification accuracy with the test data from specimen 19C-358-2, which had four SCs mirroring the length and relative depth of those found in 19C-358-1. The result showed a high TPR of 88%, coupled with a low FPR of 0.05. This outcome demonstrated that the ML model's efficacy is notably enhanced when the test data closely parallels the characteristics of the training dataset. Moreover, a similar trend of elevated TPR was observed for the four SCs within the 02-24-15 specimen when trained on the 19C-358-1 dataset, despite disparities in the height dimension, with the majority of the SCs in 02-24-15 having a smaller height compared with their counterparts in 19C-358-1. It should be noted that the first SC on 02-24-15 exhibited the lowest TPR of 0.72. This low TPR value was likely due to the small height (7.5% of wall thickness) of this particular SC.

A similar outcome was seen, with a low TPR (54%), from data corresponding to the first flaw within the 19C-358-2 specimen. The test results of this specimen were plotted in Figure 6-2, along with the original ultrasonic scan data. The orange area represents the B-scans classified as flaws, with the grey area showing the B-scans that were classified as non-flaws. The bold lines in the figure are the true positions of the flaws. Similar images for other test results can be found in Appendix B. In Figure 6-2, the first flaw is inside the weldment, and the related flaw signal was weaker than the other flaw signals in the data.

These results on SC classification seem to imply that a model trained with larger flaws can still be used for classifying data from smaller-sized flaws, achieving a reasonable TPR. However, the ML classification performance was also affected by the flaw size and flaw position in the test data (lower classification accuracy for small-sized flaws or for flaws closer to the weld centerline).



(a) 19C-358-2 (4 SC) (b) 322-14-01P (3 TFC) (c) 02-24-15 (3 TFC & 4 SC)

Figure 6-1. Confusion matrices using the model trained by 19C-358-1.

Table 6-1. Testing results using 19C-358-1 (4 SC) as the training data

Test Data Specimen (Flaws)	Accuracy	TPR	FPR	TPR for Each Flaw
19C-358-2 (4 SC)	0.88	0.83	0.05	0.54, 0.94, 0.94, 0.91
322-14-01P (3 TFC)	0.76	0.31	0.08	0.16, 0.54, 0.49
02-24-15 (3 TFC & 4 SC)	0.88	0.89	0.14	TFC (1, 0.83, 0.84) SC (0.72, 0.89, 0.95, 1)

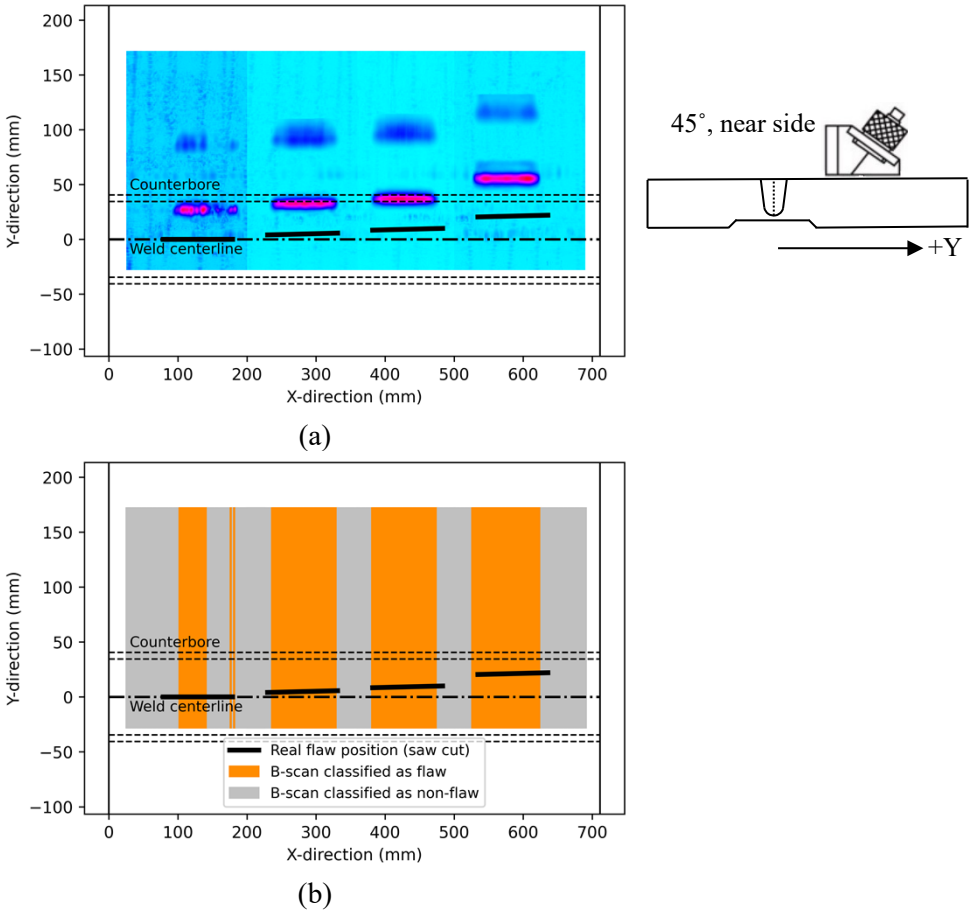


Figure 6-2. (a) Original ultrasonic scanning projection image on the top surface with a side view of inspection setup, (b) Testing results of 19C-358-2 using the model trained by 19C-358-1 overlapped on the top surface

6.1.2 Impact of Flaw Type (SC vs TFC)

The TPR for the three TFCs of 02-24-15 was seen to be high, indicating that training an ML algorithm with data from SCs (idealized flaws) may result in achieving reasonably good generalization on TFCs. While additional analyses are needed to ascertain the similarity of SC data with measurements from TFCs, these initial results seem to indicate that data from SCs contain some signal characteristics in common with data from TFCs. Given the potential for CNNs to learn localized features from the data, it is possible that a CNN may pick up on such common characteristics.

However, the good generalization performance on TFCs from 02-24-15 did not carry over to the TFCs from 322-14-01P. Specifically, the performance on the three TFCs of 322-14-01P was poor with low TPR. An examination of the data and drawings indicated that the flaws in 322-14-01P, while in the weld region, were on the far side of the weld from the ultrasonic inspection probe. This result is consistent with the studies in the report [24] that a low flaw detection rate was observed for far-side inspections. In addition, the flaw lengths and heights were relatively small. The resulting impact of the weld region and the small flaw sizes on the inspection data (in the form of poorer signal levels, increased attenuation due to the weld region, mode conversion, and noise) may be the reason for this poor performance on 322-14-01P.

Collectively, these results appear to confirm the assumption that the fundamental differences in the data due to the type of flaws can influence the ML classification accuracy.

6.1.3 False Positives and False Negatives

The results shown above, in most cases, indicated a relatively low FPR. However, the FPR in the case of 02-24-15 was higher (0.14), likely due to the impact of the weld region on the inspection data from flaws close to the weld center line. It is worth noting that, in addition to the TPR and FPR, the false negative rates (FN) are equally important, as the FN reflects the potential for missing flaw signals. All else being equal, a model with a lower FN (ideally zero) would be preferred. Results on the stainless steel weldments appeared to show moderate FNR (0.16 and 0.1 for specimens 19C-358-2 and 02-24-15), with a higher FNR when dealing with flaws on the far side of the weld center line (322-14-01P). Again, it is likely that the impact of the weld microstructure on the data, and the resulting differences between this data and the training dataset, may have impacted the FN along with the low TPR on data from this specimen.

6.1.4 Preprocessing Stages

Changing the preprocessing approach on test data, compared to the preprocessing methods used on the training data, is expected to have a negative impact on the ML performance. In part, this is because of the changes in the signal characteristics that are imparted by the preprocessing methods (cropping, gain adjustment, normalization, etc.).

To study the effect of the preprocessing procedures on the ML performance, the model trained with 19C-258-1 was again tested with the data from the other three specimens (19C-258-2, 02-24-15, and 322-14-01P). However, the B-scan images from the three test specimens were preprocessed using a different preprocessing method compared to the data used for training, with the goal of demonstrating the impact of varying the data preprocessing stages on the ML performance. First, the B-scan images without gain adjustment but with image cropping were tested using the model. Then, the B-scan images without image crop but with gain adjustment were tested. These results were compared with the results in Table 6-1. Table 6-2 presents the test results on the B-scan images without gain adjustment. Compared with Table 6-1, the performance was dramatically decreased on the data of specimens 19C-358-2 and 322-14-01P with lower TPRs for each flaw. For specimen 02-24-15, the classification performance was slightly improved using B-scans without gain adjustment. This could be caused by the high amplitudes of the flaw features in the original B-scan images of these specimens. Therefore, the gain adjustment did not appear to benefit the classification performance on this specimen. Overall, the gain adjustment could improve the classification performance of the ML model, especially for data with low amplitudes of flaw features.

The test results on the B-scan images without image crop are presented in Table 6-3. Compared with the results in Table 6-1, the results without image cropping were poorer, with low or even zero TPRs for each flaw. Because image cropping may be used to remove noise due to coupling variations and any back-surface reflections, the B-scan images after image cropping became much cleaner, which may lead to better classification performance. Additionally, after cropping and downsampling, the flaw features are exaggerated, which could also improve the performance. Therefore, image cropping may enhance the signal in the B-scan images and is beneficial to improving the classification performance.

Table 6-2. Testing results using B-scan images without gain adjustment.

Test Data Specimen (Flaws)	Accuracy	TPR	FPR	TPR for Each Flaw
19C-358-2 (4 SC)	0.58	0.28	0	0, 0.1, 0.17, 0.86
322-14-01P (3 TFC)	0.77	0.15	0	0, 0, 0.42
02-24-15 (3 TFC & 4 SC)	0.87	0.94	0.14	TFC (1, 0.93, 0.97) SC (0.78, 0.90, 1, 1)

Table 6-3. Testing results using B-scan images without cropping.

Test Data Specimen (Flaws)	Accuracy	TPR	FPR	TPR for Each Flaw
19C-358-2 (4 SC)	0.42	0	0	0, 0, 0, 0
322-14-01P (3 TFC)	0.73	0	0	0, 0, 0
02-24-15 (3 TFC & 4 SC)	0.71	0.40	0.08	TFC (1, 0.8, 0) SC (0.66, 0.88, 0, 0)

6.1.5 Impact of Training Data Size

Data from specimens 322-14-01P and 02-24-15 were tested using a model trained on data from specimens 19C-358-1 and 19C-358-2. In this scenario, the training dataset size was doubled compared with the model used in Table 6-1. The results for 02-24-15 remained relatively consistent in terms of TPR and FPR. In the case of 322-14-01P, there appeared to be some improvement; however, the FPR remained elevated. Consequently, increasing the volume of training data did not yield a substantial improvement in the testing performance for 322-14-01P and 02-24-15 (refer to Table 6-4 for details).

It is important to set these results in context. In general, for DNNs, training and classification performance tends to improve with more data. However, the increase in data needs to be representative. In this case, the increase in data roughly doubled the number of available B-scan images for training. However, these were all from SCs that were fairly similar in terms of length, through-wall height, and placement relative to the weld center line. However, the test data consisted of a combination of SC and TFC, with placement on both the near and far sides of the weld center line. Thus, the numeric increase in training data was not accompanied by an increase in data diversity that captured the expected variation in flaw type, location, and size in the test data. The result was no significant improvement in testing performance.

Table 6-4. Testing results using 19C-358-1 and 19C-358-2 as the training data.

Test Data Specimen (Flaws)	Accuracy	TPR	FPR	TPR for Each Flaw
322-14-01P (3 TFC)	0.62	0.95	0.5	0.91, 0.92, 1
02-24-15 (3 TFC & 4 SC)	0.87	0.94	0.19	TFC (1, 1, 1) SC (0.75, 0.91, 0.98, 1)

6.1.6 Training with TFCs

Table 6-5 summarizes the accuracy, TPR, and FPR, using the flaw data from TFCs in specimen 322-1401P as the training dataset. Unfortunately, when employing the three TFCs of 322-14-01P as the training data, the overall testing performance on the other three specimens yielded unsatisfactory results. The test outcomes for specimens 19C-358-1 and 19C-358-2 exhibited a discernible bias toward flaws, with most non-flaws being incorrectly classified as flaws. Despite the elevated TPR values, this type of bias increases the FPR. In contrast, the trained model exhibited a different type of bias when applied to 02-24-15, where it missed all flaws. It is not clear if this is due to a significant difference in TFC signatures across these specimens (perhaps due to differences in flaw characteristics in these two specimens), or due to improper training (for instance, overfitting of the training data). In either case, the results demonstrate the need for care in choosing the training data and in monitoring the training and validation loss curves for overfitting during the training process.

A second model was trained on the three TFCs of 02-24-15. Poor classification performance was observed across specimens 19C-358-1, 19C-358-2, and 322-14-01P, as seen in the last four rows of Table 6-5. Notably, high TPRs were only achieved for the four SCs of 02-24-15, as evident in the last row of Table 6-5. One plausible explanation for this phenomenon is the close association between the three TFCs and the four SCs within the same specimen. It is likely that these two datasets share certain consistent ultrasonic features within the B-scan images. However, this shared coherence did not translate into effective generalization when applied to other SC data, as the model's performance remained suboptimal.

These results also indicate best practice may dictate data from a single specimen not be split among training and test data, especially if the goal is to perform an unbiased evaluation of the ML algorithms.

Table 6-5. Test results using TFCs as training data.

Test Data Specimen (Flaws)	Accuracy	TPR	FPR	TPR for Each Flaw
1. 322-14-01P (3 TFC) as training				
19C-358-1(4 SC)	0.66	0.88	0.67	0.56, 1, 0.97, 1
19C-358-2 (4 SC)	0.88	1	0.99	1, 1, 1, 1
02-24-15 (3 TFC+ 4 SC)	0.60	0	0	TFCs (0, 0, 0, 0) SCs (0, 0, 0, 0)
2. 02-24-15 (3 TFC) as training				
19C-358-1(4 SC)	0.43	0.02	0	0, 0.08, 0, 0
19C-358-2 (4 SC)	0.53	0.20	0	0, 0, 0, 0.82
322-14-01P (3 TFC)	0.77	0	0	0, 0, 0
02-24-15 (4 SC)	0.84	0.92	0.11	0.72, 0.89, 1, 1

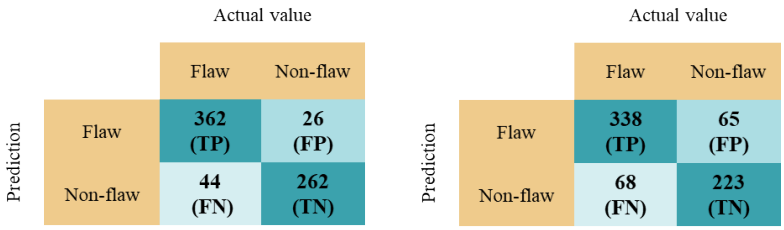
6.1.7 Increasing Training Data Diversity through Transfer Learning

Recognizing the poor classification performance of the model trained on data from 322-14-01P when applied to specimens 19C-358-1 and 19C-358-2, the model was retrained using the data extracted from the four SCs of 02-24-15. The B-scans from these four SCs displayed flaw signatures similar to the B-scan images of specimens 19C-358-1 and 19C-358-2. The test performance of this retrained model on specimens 19C-358-1 and 19C-358-2 is visually depicted in Figure 6-3 and quantitatively summarized in Table 6-6. The classification performance has a remarkable improvement compared to the results reported in Table 6-5. To further assess and compare the efficacy of different models, four ROC curves were

generated, as illustrated in Figure 6-4. These ROC curves correspond to the test results on specimens 19C-358-1 and 19C-358-2, using four models:

(1) a model originally trained with 322-14-01P,
(2) a model trained exclusively with data from 02-24-15,
(3) a model trained using the four SCs of 02-24-15 and
(4) a model initially trained with 322-14-01P but subsequently retrained with data from the four SCs of 02-24-15.

The retrained model showed the highest level of performance among all models, as evident from Figure 6-4. This superiority can be attributed to the diverse dataset used for the retraining, encompassing both TFC and SC data. Moreover, the incorporation of the four SCs, which closely resembled specimens 19C-358-1 and 19C-358-2, contributed to the model's enhanced capability. Consequently, this retrained model has a superior performance over both the model trained solely on 322-14-01P and the model relying on data from the four SCs of 02-24-15. These results also demonstrated that a retraining procedure, using a trained network as the starting point when adding more data (a CNN trained with SCs subsequently retrained to include only TFCs), may be helpful in improving classification performance. This type of transfer learning may also be useful in improving the performance of a previously qualified CNN with site-specific data, though the retrained model will likely need to be re-qualified.



(a) 19C-358-1 (4 SC) (b) 19C-358-2 (4 SC)

Figure 6-3 Confusion matrices using the model first trained by 322-14-01P and then retrained by 02-24-15 (four SC).

Table 6-6. Test results using the model that was first trained with 322-14-01P and retrained with four SCs of 02-24-15.

Test Data Specimen (Flaws)	Accuracy	TPR	FPR	TPR for each Flaw
A (4 SC)	0.90	0.89	0.09	0.92, 99, 0.93, 0.72
B (4 SC)	0.81	0.83	0.22	0.46, 0.92, 0.95, 1

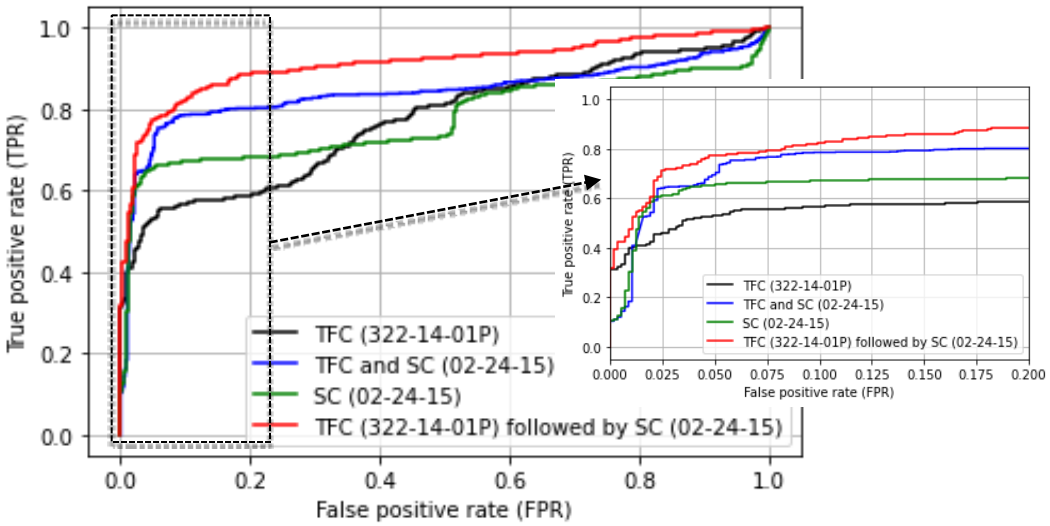


Figure 6-4. ROC curves of four training conditions and tested with specimens 19C-358-1 and 19C-358-2.

6.1.8 Testing Using Different Probes

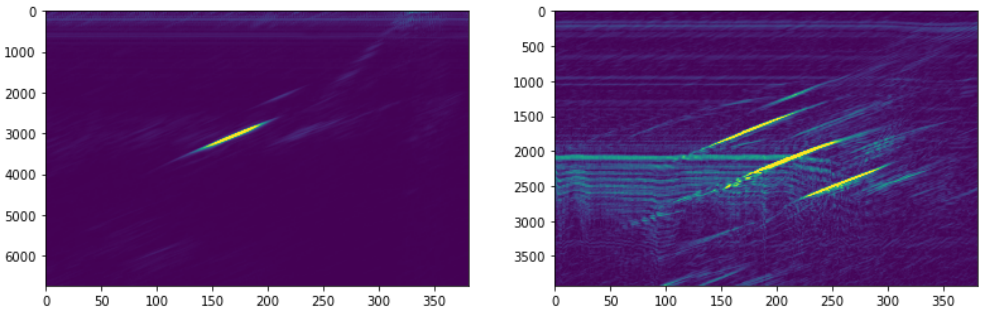
A CNN model was trained on shear probe data and subsequently tested on longitudinal probe data, or vice versa, to assess its ability to generalize across different probe types. In this section, the CNN model was first trained on the 19C-358-1 dataset, acquired using the SwRI 45˚ shear probe, and its performance evaluated using data from various specimens, all acquired using the longitudinal GEIT 45˚ probe.

Table 6-7 summarizes the test results for four specimens. The classification performance proved subpar across the board, with low accuracies and, remarkably, zero true positive rates (TPRs) for specimens 19C-358-2, 322-14-01P, and 02-24-15. Only when the same specimen, 19C-358-1, was involved were non-zero TPRs observed for the initial three flaws on that specimen. These findings, not unexpected, indicate that the ML model trained with the shear probe data exhibits inadequate generalization capabilities when applied to data collected with the longitudinal probe, even when working with data from the same specimen.

To illustrate this discrepancy further, Figure 6-5 displays the original B-scan images collected for the same flaw in 19C-358-1. Figure 6-5 (a) showcases the B-scan image obtained using the shear probe SwRI, revealing a pattern vastly different from the image captured using the longitudinal probe GEIT, as seen in Figure 6-5 (b). The B-scan image obtained with the shear probe had relatively low noise, with only one flaw feature in the center of the image. The B-scan image obtained via the longitudinal probe exhibits a higher level of noise and incorporates features related to mode conversion (L→S). Consequently, the ML model, trained primarily on shear probe data, exhibited suboptimal generalization performance when confronted with data collected using the longitudinal probe.

Table 6-7. Test results using the model trained with 19C-358-1 SwRI 45˚

Test Data Specimen, Probe, Angle	Accuracy	TPR	FPR	TPR for Each Flaw
19C-358-1 GEIT 45˚	0.62	0.35	0	0.7, 0.55, 0.14, 0
19C-358-2 GEIT 45 ˚	0.42	0	0	0, 0, 0, 0
322-14-01P GEIT 45 ˚	0.73	0	0	0, 0, 0
02-24-15 GEIT 45 ˚	0.59	0	0	SC (0, 0, 0, 0) TFC (0, 0, 0)



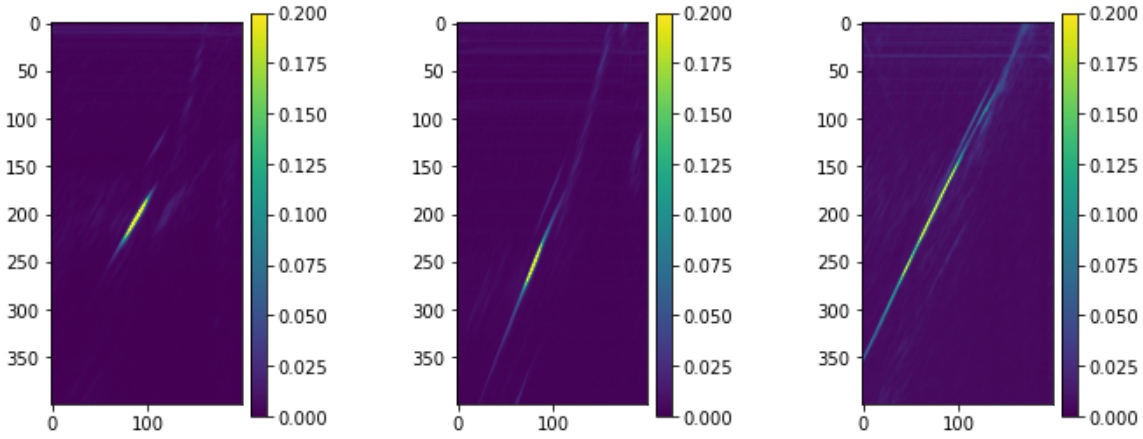
(a) Shear probe (SwRI 45 ˚) (b) Longitudinal probe (GEIT 45 ˚)

Figure 6-5. B-scan images collected on flaw 1 of 19C-358-1 using different probes.

6.1.9 Testing Using Different Refracted Angles

A CNN model was trained using data collected at one specific refracted angle and subsequently tested with data from other refracted angles. The data employed in this analysis were exclusively derived from the shear probe, focusing on specimen 19C-358-1. These data encompassed three refracted angles: 45˚, 60˚, and 70˚. Figure 6-6 visually presents the original B-scan images associated with these three angles, each manifesting a unique pattern. Notably, the 45˚ image exhibited the smallest flaw features, whereas the 70˚ image displayed significantly longer flaw features.

Table 6-8 summarizes the test results derived from the models trained with data from one angle at a time. Unfortunately, across all three models, the testing performance was characterized by either low true positive rates (TPR) or high false positive rates (FPR). The solitary exception emerged with the model trained on the 70˚ data, which yielded a relatively promising classification outcome when tested on 45˚ data, with a high TPR (0.80) and a low FPR (0.04). The reason for this exception is still unclear and needs more investigation. These findings underscore the challenge of achieving generalization when training a CNN model on data from one specific refracted angle for application on data acquired at a different refracted angle. Improving performance is likely to require training the ML to learn from data from multiple refracted angles, either by including the data in the training data set up-front or through a retraining process.



(a) Refracted angle 45 ˚ (b) Refracted angle 60 ˚ (c) Refracted angle 70 ˚

Figure 6-6. B-scan images collected on flaw 1 of 19C-358-1 using different refracted angles.

Table 6-8. Test results using the model trained with 19C-358-1 SwRI 45˚, 60˚, and 70˚

Testing / Training	SwRI 45˚	SwRI 60˚	SwRI 70˚
SwRI 45˚	N/A	Acc.=0.42, TPR=0, FPR=0	Acc.=0.415, TPR=0, FPR=0
SwRI 60˚	Acc.=0.65, TPR=0.99, FPR=0.84	N/A	Acc.=0.43, TPR=0.02, FPR=0
SwRI 70˚	Acc.=0.86, TPR=0.80, FPR=0.04	Acc.=0.60, TPR=1, FPR=0.97	N/A

6.2 TRAINING AND TESTING USING DMW AND STAINLESS-STEEL SPECIMENS

In prior sections, the CNN model was exclusively trained on stainless-steel weld data and subsequently tested on data from other stainless-steel weldments. In this section, the performance of the ML model was studied when the model was trained on one type of specimen data and tested on a different type of specimen data, specifically DMW specimens.

Two models were trained with stainless-steel specimens: one trained on the 19C-358-1 SwRI 45˚ dataset and another trained on the 02-24-15 SwRI 45˚ dataset. These models were used to assess their performance on data originating from two DMW specimens, 8C-032 and 8C-091, both captured using the SwRI 45˚ probe. The results of this evaluation are presented in Table 6-9. Both models exhibited suboptimal classification performance when applied to the DMW specimens. While the classification accuracy for 8C-032 was relatively high in both cases (0.85 and 0.87), the TPRs remained notably low (0.07 and 0.39). Indeed, the high classification accuracy appears to be the result of a larger no-flaw region in the specimens and a very low FPR. Furthermore, the TPRs for 8C-032 displayed mixed results, with the highest TPR reaching 0.77 for the fourth flaw in the model trained by 02-24-15, while the maximum overall TPR achieved by the model trained by 19C-358-1 was only 0.36. The performance on the 8C-091 dataset was unsatisfactory, with almost zero TPRs for each flaw. Consequently, models trained with stainless-steel specimen data struggled to classify flaws within the DMW specimens effectively. These outcomes underscore again that models need to be trained on a diverse dataset for improved performance.

Table 6-9. Test results on the DMW specimens using the model trained with stainless steel specimens

Test Data Specimen (Flaws)	Accuracy	TPR	FPR	TPR for Each Flaw
1. 19C-358-1 (4 SC) as training				
8C-032 (4 TFC)	0.85	0.07	0.02	0, 0, 0, 0.36
8C-091 (2 notches & 2 TFC)	0.57	0.02	0	0.07, 0, 0, 0
2. 02-24-15 (4 SC and 3 TFC) as training				
8C-032 (4 TFC)	0.89	0.39	0.02	0, 0.45, 0.37, 0.77
8C-091 (2 notches & 2 TFC)	0.56	0	0	0, 0, 0, 0

When employing DMW specimens as the training dataset, two models were developed using the data extracted from specimens 8C-031 and 8C-091, both employing the SwRI 45˚ probe. Subsequently, these two models were tested using data obtained from stainless-steel specimens 19C-358-1 and 02-24-15. The outcomes of these tests are summarized in Table 6-9. Much like the results detailed in Table 6-9, the analysis revealed that both models exhibited low accuracies and true positive rates (TPRs) when applied to the two stainless-steel weld specimens. Specifically, for the majority of flaws in the test data, the TPR was close to, or at zero. The only exception was observed for flaw #1 of specimen 19C-358-1, where the TPR reached 0.9 when employing the model trained with data from 8C-032. It is evident that the models trained using DMW specimens failed to demonstrate effective generalization when applied to the flaw classification of stainless-steel specimens featured in the study. Additional analysis using DMW data is ongoing and expected to be included in subsequent reports and publications.

Table 6-10. Test results on the stainless steel specimens using the model trained with DMW specimens

Test Data Specimen (Flaws)	Accuracy	TPR	FPR	TPR for Each Flaw
1. 8C-032 (4 TFC) as training				
19C-358-1 (4 TFC)	0.54	0.22	0.02	0.9, 0, 0, 0
02-24-15 (4 SC and 3 TFC)	0.62	0.04	0	4SC (0, 0, 0, 0.1) 3TFC (0, 0.1, 0.08)
2. 8C-091 (2 notches & 2 TFC) as training				
19C-358-1 (4 TFC)	0.27	0.03	0.39	0.06, 0, 0.03, 0.03
02-24-15 (4 SC and 3 TFC)	0.58	0.1	0.1	4SC (0.38, 0, 0, 0) 3TFC (0.5, 0.13, 0.19)

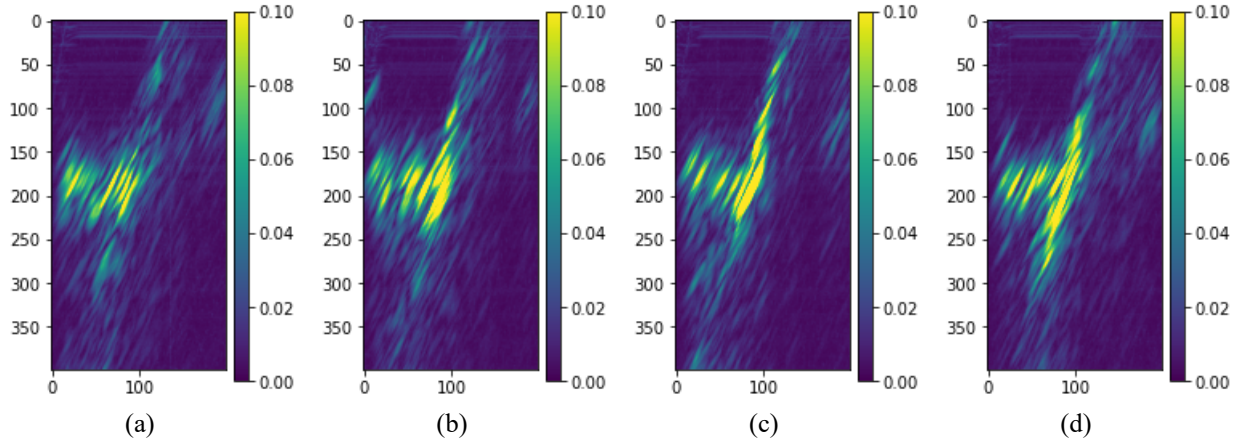


Figure 6-7. Preprocessed B-scan images (400×200) in DMW specimen 8C-032 at the center position of (a) flaw 1, (b) flaw 2, (c) flaw 3, (d) flaw.

7. SUMMARY

In the context of the nuclear industry, ML algorithms are likely to be deployed in one of several ways, with oversight assistance (assisting site personnel in reviewing results) and analyst assistance (identifying regions of interest that are then subject to human analyst review and dispositioning) likely to be the near-term options based on publicly available information. Whether applied for oversight assistance or analyst assistance, or as part of other solutions for automated analysis of NDE data, there is a need to develop the technical bases for guidance on the application of ML for NDE of nuclear power plant (NPP) components. Implicit in the development of the technical bases is the need to understand the factors that influence the performance of ML methods.

Research to date has demonstrated that, while ML has the potential for high accuracy in the context of the classification of ultrasonic NDE data, the accuracy may be impacted by several factors. The accuracy of the classification appears to vary based on the flaw type, size, location, and weld type and points to the need for representative data from a range of flaw types, sizes, and locations. The results also appear to indicate that ML may be able to learn key characteristics of flaw signals that are translatable across flaw types, though such transferability does not seem to apply to other factors (weld type, inspection mode, angle, etc.). Transfer learning and techniques for ML model fine-tuning are among the approaches that can help improve performance over time and reduce the requirements for large data sets. The classification performance was also seen to be affected by the preprocessing procedures of the B-scan images. These results are not surprising, as key image features are influenced by the preprocessing.

The results to date indicate a potential for wide variation in performance with ML algorithms capable of achieving very high or very low classification accuracy. Collectively, these results appear to indicate the following:

- A CNN or similar ML algorithm may be able to learn key features of flaws and non-flaws using data from simple flaws (SCs) and generalize well to other flaw types as long as there are no other factors (such as inspection through welds) (Section 6.1.1). Note that these factors may be accounted for by the use of an expanded training data set (Section 6.1.7), especially as ML algorithms will be expected to accommodate nominal weld geometrical variances and associated noise in B-scan images given the ASME BPVC defined target inspection volume (lower 1/3 of the volume of the weld from the ID surface).
- Generalization performance may vary depending on the flaw size and location. Smaller (both length and height) flaws tend to be more difficult to detect using ML, and the difficulty appears to increase if the flaw is in the vicinity of a weld. Such an effect may also be applicable to the manual (human) analysis of the data and may indicate inherent challenges in the data itself (Sections 6.1.1 and 6.1.2).
- A high TPR may not, by itself, be indicative of good classification accuracy. In many instances, the FPR was equally high and indicative of a CNN that appeared to be biased toward flaws, with most non-flaws misclassified as flaws (Section 6.1.3).
- A retraining procedure, using a trained network as the starting point when adding more data (a CNN trained with SCs subsequently retrained to include only TFCs), may be helpful in improving classification performance. This type of transfer learning may be, for instance, useful in improving the performance of a qualified CNN with site-specific data (Section 6.1.7).
- Data sets used for training should be diverse and representative of the types of data expected to be encountered during use (testing) (Sections 6.1.2, 6.1.8, and 6.1.9). For example, if the ML is intended to be applied to DMW specimens, then the training data should contain data from similar specimens and flaw types. In addition, the training data should reflect the expected diversity in flaw types and locations to ensure the ML training is adequate. It should be noted that

variations in inspection angles and probe frequencies are likely if the inspection procedures are changed; as a result, and as depicted in the initial results in Section 6.1.8, ML models are unlikely to be sufficiently accurate when applied to data collected using a different procedure.

- Preprocessing procedures should be consistent across the data used for training and testing ML. Inconsistencies in these stages can have a major impact on classification accuracy (Section 6.1.4). Included is the need to have consistency in labeling the training data for supervised ML methods (Section 4.3).
- The analyses did not explicitly quantify the uncertainty in the ML classification results. It is worth noting that the CNN (or any ML) will provide a result when given an input. While comparisons of the classification to a known class can be made in controlled data sets, such comparisons are difficult in a field setting where the true class is not known *a priori*. As a result, approaches that estimate confidence in the classification result will be helpful in interpreting the classification outputs from ML models.

The variability in ML performance appears to be driven by the data used for training (the type and potentially quality) and the ML architecture. These findings are anticipated to have broad applicability to ML algorithms as suggested by comparative assessments in the ML research community. The comparative assessments demonstrate the influence of these factors on classification accuracy across multiple ML algorithms. The findings also indicate that, in the context of the nuclear industry, applications of ML for the classification of NDE data may require substantial care to ensure that the algorithms are well-trained to avoid over-fitting (memorization) and validated using an extensive data set to enhance confidence in the result.

The findings to date seem to indicate that, while the application of ML for classification can benefit NDE in nuclear power applications, care must be taken with the selection of data sets for training, and a comprehensive assessment of the classification performance is necessary. Inherent in this comprehensive assessment is the evaluation of TPR, FPR, and FN, as well as the tradeoffs between these quantities. For the purposes of this report, basic criteria on what constitutes good classification performance were used. However, qualification standards for ML for NDE applications may differ, and the applicability of these techniques for field examinations will need to be determined in concert with qualification criteria.

Given the results seen to date, likely requirements that need to be established before ML is qualified for NDE data analysis include the following:

- Establishing confidence in the ML results: training, validation, and test data requirements and distributional information will need to be specified. A large common data set, with some portion of the data made available for training and validation, will be required based on the distribution specifications. The data set should include data from the procedure(s) of interest and incorporate data from a variety of flaw sizes and locations. The selected procedures are expected to limit the range of frequencies and probes used to collect the data.
- Establishing confidence in the ML results: test data requirements will need to be specified. Test data should generally be selected from specimens that do not contribute to the training data sets. This separation avoids information leakage, limits the potential for artificially biasing the ML performance, and allows a better evaluation of the ML algorithm.
- Establishing confidence in the ML results: qualification. Metrics for qualification are likely going to be similar to existing qualification requirements for personnel and procedures, though analysis guidelines for ML will need to be developed and incorporated into procedures. Performance requirements across multiple metrics are likely to be necessary, and at a minimum, performance requirements for TPR, FPR, and FNR (false negative rates) may be needed, though ROC curves or similar quantities can allow a better assessment of the tradeoffs associated with the trained ML

algorithm. Note that these minimum performance targets may be adapted from similar targets if they are specified for manual analysis.

- Documentation: ML parameters (weights and architecture) as well as the hyperparameters used in the training process, will need to be documented. Performance is a function of the model and specific parameters (weights) and model-specific hyperparameters (such as model structure, learning rates, and learning optimizer parameters), and at least some of these will need to be included as essential variables in any performance demonstration. In addition, information about the software frameworks, data sets, and inspection procedures may need to be defined as part of the essential variables associated with the ML algorithm. Note that additional information, such as procedure-specific information used to generate the training data, may also need to be captured and documented with the ML model to ensure greater confidence in the ML results when applied to a test data set.
- Establishing confidence in the ML results: qualification. Approaches to qualification (personnel vs procedure) need to be determined. If personnel qualification is determined to be the selected pathway, variability because of the use of different procedures will need to be handled. Requalification and potential site-specific qualification will need to be determined, especially if the site-specific qualification will require retraining. A specific question would be defining the threshold for triggering a requalification if retraining occurs. Requirements for qualification (possibly performance demonstration) will also need to be defined. Qualification requirements are likely to vary based on the deployment use cases. For instance, the use of ML to assist or supplement a qualified inspector will likely require the ML to detect all flaws while maintaining a reasonable rate of additional calls, with the expectation that the qualified analyst will disposition all indications that are flagged by the ML. On the other hand, using the ML in a fully automated analysis mode will require a high true positive rate (low false negative rate) and a low false call rate.
- Uncertainty quantification of the ML output will be important, though this is still a topic of research within the ML community and methods are relatively limited at present. The quantification of uncertainty in ML results will likely be more important as advances in algorithms allow their use for classification independent of human analysts.
- Verification and validation of the software stack used for deployment. Current ML algorithms use either custom-developed software for training and testing (inference) or use open-source frameworks such as Tensorflow and Pytorch (two of the more popular frameworks). In either case, the rapid pace of development of ML technology requires verifying that the implementations are correct and provide the expected results on benchmark data sets. Complicating the V&V process is the difficulty in achieving identical results by repeating the ML training process without taking additional steps, such as using the same initial values for the ML parameters and ensuring that the training data is used in the same sequence each time. While the use of open-source platforms, reference data sets, and a qualification procedure may help mitigate concerns relative to the V&V of the software stack, this is likely to be a topic that needs further investigation.

While a comprehensive assessment of codes and standards is needed, it is already apparent through other research [6] that there are no consensus standards yet available on the application of AI/ML for nuclear power or in any other applications, though several standards organizations are working on developing such standards. It is likely that, while existing requirements for ultrasonic inspection may be directly applicable to ML, changes in the BPV Code or guidance for documenting the ML and its performance will be needed.

Ongoing research is continuing as the evaluation of ML incorporates additional data sets, including the use of data augmentation techniques and the incorporation of simulation data with empirically derived

data for training the algorithms. Research is also assessing the robustness and interpretability of ML models for NDE data analysis utilizing methods proposed in the machine learning literature. The goal of these analyses is to evaluate methods for assessing the sensitivity of the models to noise in the data, determine if the models are overfitting the training data, and determine whether a reasonable level of interpretability is possible. Collectively, these assessments are expected to be useful in increasing confidence in the robustness of the results. Applications beyond simple classification (including image segmentation) are also being assessed, and the findings are being used to develop recommendations for validation and qualification of ML prior to field use.

8. REFERENCES

1. "An Approach For Using Probabilistic Risk Assessment In Risk-Informed Decisions On Plant Specific Changes To The Licensing Basis," Nuclear Regulatory Commission (2011).
2. B. Bishop et al., "Materials Reliability Program: Risk-Informed Revision of ASME Section XI Appendix G - Proof of Concept (MRP-143)," 1009546, Electric Power Research Institute (2005).
3. L. Udpa and S. Udpa, "Eddy Current Defect Characterization Using Neural Networks: Materials Evaluation," *NDT International* **23** (6), pp.358, Elsevier (1990).
4. M. Dennis et al., "Artificial Intelligence Strategic Plan, Fiscal Years 2023-2027," NUREG-2261, Nuclear Regulatory Commission (NRC) (2023).
5. H. Sun, P. Ramuhalli, and R. E. Jacob, "Machine Learning For Ultrasonic Nondestructive Examination of Welding Defects: A Systematic Review," *Ultrasonics* **127**, pp.106854 (2023).
6. M. Muhlheim et al., "Status Report on Regulatory Criteria Applicable to the Use of Artificial Intelligence (AI) and Machine Learning (ML)," ORNL/SPR-2023/3072, Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States) (2023).
7. I. Virkkunen et al., "Augmented Ultrasonic Data for Machine Learning," *Journal of Nondestructive Evaluation* **40** (1), pp.4 (2021).
8. C. R. Farrar and K. Worden, *Structural Health Monitoring: A Machine Learning Perspective*, John Wiley & Sons (2012).
9. F.-G. Yuan et al., "Machine Learning For Structural Health Monitoring: Challenges And Opportunities," in *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2020* **11379**, p. 1137903, SPIE (2020).
10. M. Flah et al., "Machine Learning Algorithms in Civil Structural Health Monitoring: A Systematic Review," *Archives of Computational Methods in Engineering* **28** (4), pp.2621–2643 (2021).
11. U. Dackermann, B. Skinner, and J. Li, "Guided Wave–Based Condition Assessment of In Situ Timber Utility Poles Using Machine Learning Algorithms," *Structural Health Monitoring* **13** (4), pp.374–388, SAGE Publications (2014).
12. Y. Liu and Y. Bao, "Review on Automated Condition Assessment of Pipelines With Machine Learning," *Advanced Engineering Informatics* **53**, pp.101687 (2022).
13. J. L. Rose, *Ultrasonic Waves in Solid Media*, Acoustical Society of America (2000).
14. ASNT, "Nondestructive Testing Handbook, Third Edition: Volume 7, Ultrasonic Testing," American Society for Nondestructive Testing, Columbus, Ohio (2007).
15. L. Schmerr and J.-S. Song, "Ultrasonic Nondestructive Evaluation Systems: Models and Measurements," Springer US (2007).
16. M. D. Wilkinson et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship," 1, *Scientific Data* **3** (1), pp.160018, Nature Publishing Group (2016).
17. T. Koskinen et al., "The Effect of Different Flaw Data to Machine Learning Powered Ultrasonic Inspection," *Journal of Nondestructive Evaluation* **40** (1), pp.24 (2021).

18. N. Munir et al., “Investigation of Deep Neural Network With Drop Out for Ultrasonic Flaw Classification in Weldments,” *Journal of Mechanical Science and Technology* **32** (7), pp.3073–3080 (2018).
19. N. Munir et al., “Performance Enhancement of Convolutional Neural Network for Ultrasonic Flaw Classification by Adopting Autoencoder,” *NDT & E International* **111**, pp.102218 (2020).
20. Z. Chen et al., “Automatic Recognition of Weld Defects in TOFD D-Scan Images Based on Faster R-CNN,” *Journal of Testing and Evaluation* **48** (2), pp.811–824, ASTM International (2018).
21. Y. Yan et al., “A Deep Learning-Based Ultrasonic Pattern Recognition Method for Inspecting Girth Weld Cracking of Gas Pipeline,” *IEEE Sensors Journal* **20** (14), pp.7997–8006 (2020).
22. C. Garbin, X. Zhu, and O. Marques, “Dropout vs. Batch Normalization: An Empirical Study of Their Impact to Deep Learning,” *Multimedia Tools and Applications* **79** (19), pp.12777–12815 (2020).
23. A. F. Agarap, “Deep Learning using Rectified Linear Units (ReLU),” arXiv:1803.08375, arXiv (2019).
24. J. Harrison et al., “Evaluating Flaw Detectability Under Limited-Coverage Conditions,” PNNL-30238, Pacific Northwest National Laboratory, pp. 1–102 (2020).

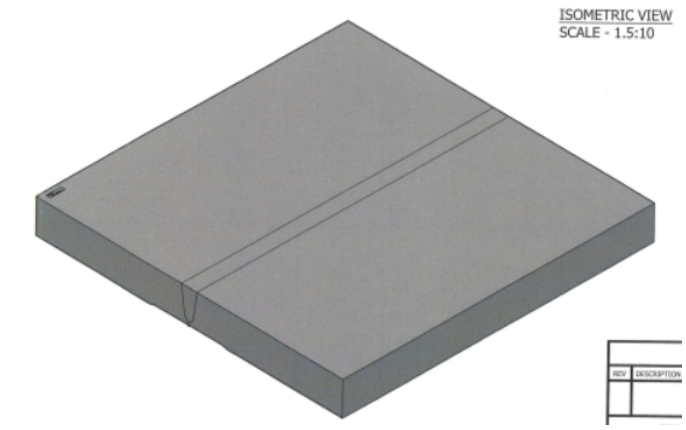
APPENDIX A. SPECIMEN AND FLAW INFORMATION

APPENDIX A. SPECIMEN AND FLAW INFORMATION

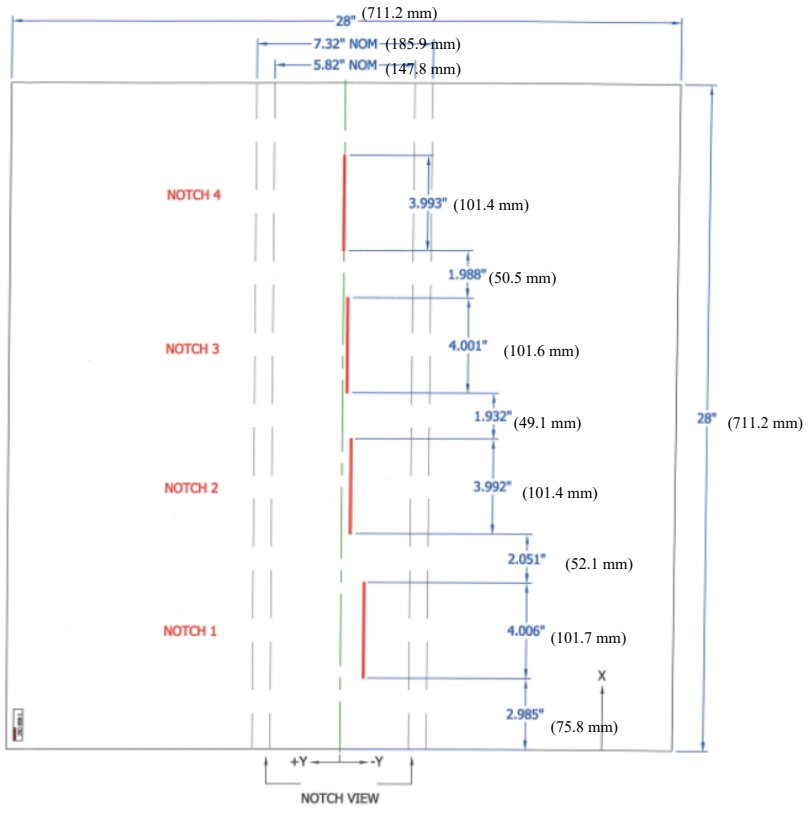
The detailed information on the specimens and flaws is summarized in the table below.

Table A-1. Summary of the specimen information

Specimen #	Description	Base Material	Material Class	Weld Material	Flaws	Geometry	Length (mm)	Width (mm)	Thickness (mm)	ID (mm)	OD (mm)
19C-358-1	Custom SS Plate	304	SS	308 SS	SC	Plate	711.2	711.2	76.2		
19C-358-2	Custom SS Plate	304	SS	308 SS	SC	Plate	711.2	711.2	31.8		
322-14-01P	14 in. Pipe to CSS Valve	316	SS	316 SS	TFC	Pipe Section	340.4		38.6	292.1	369.3
		SA351	CSS								
02-24-15	24 in. Sch80 304 SS	A358	SS		TFC, SC	Pipe	508.0		35.3	539.0	609.6
8C-032	DMW specimen	A321	CS	309 SS	TFC	Pipe	256.1		28.6	266.7	323.9
		316	SS	Inconel 182		Safe End			44.5	260.4	349.3
		SA508	CS			Nozzle			38.1	273.1	349.3
8C-091	DMW PZR Surge Nozzle Specimen		CSS		EDM, TFC	Pipe	608.1		39.4	244.9	323.7
			CS			Nozzle					

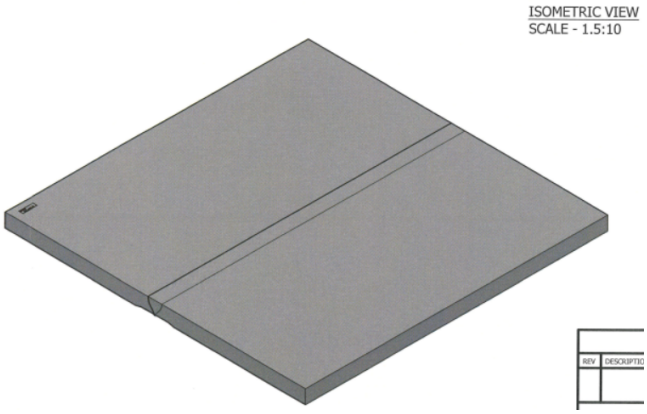


(a)

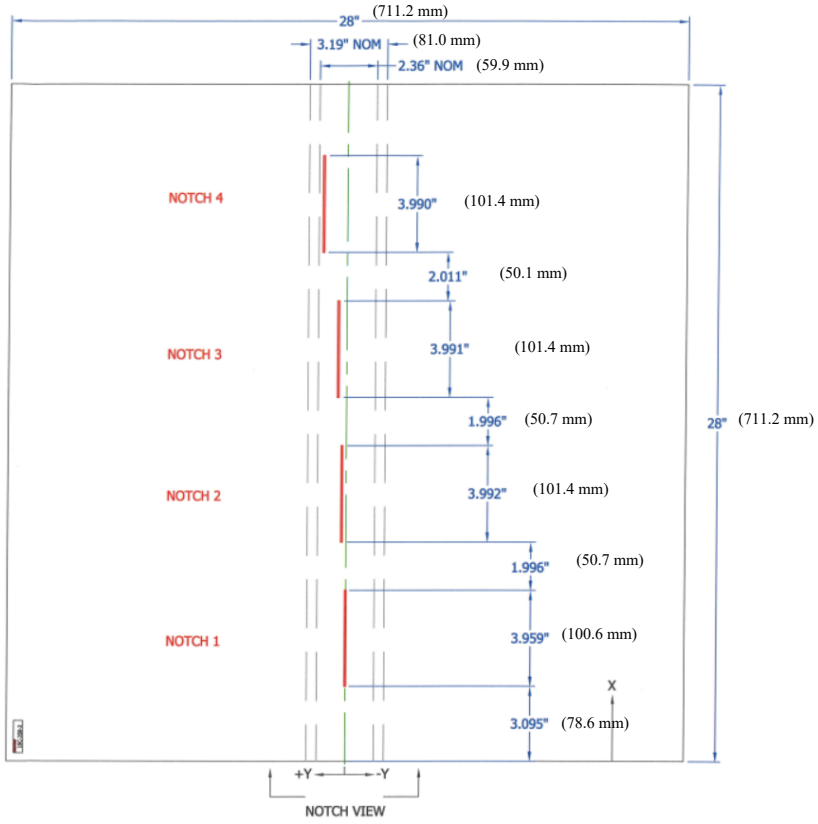


(b)

Figure A-1. (a) Specimen 19C-358-1 and (b) flaw positions on the specimen 19C-358-1.



(a)



(b)

Figure A-2. (a) Specimen 19C-358-2 and (b) flaw positions on the specimen 19C-358-2.

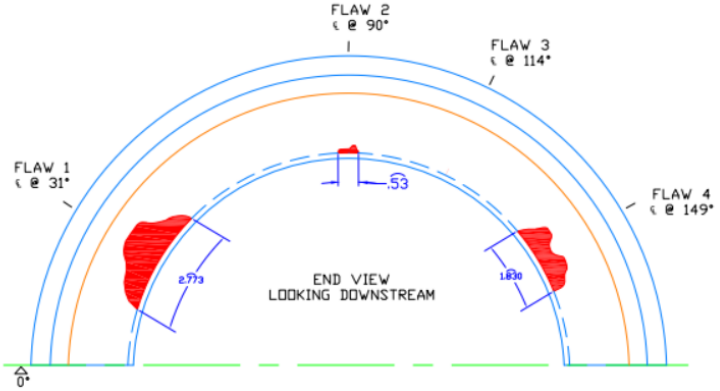
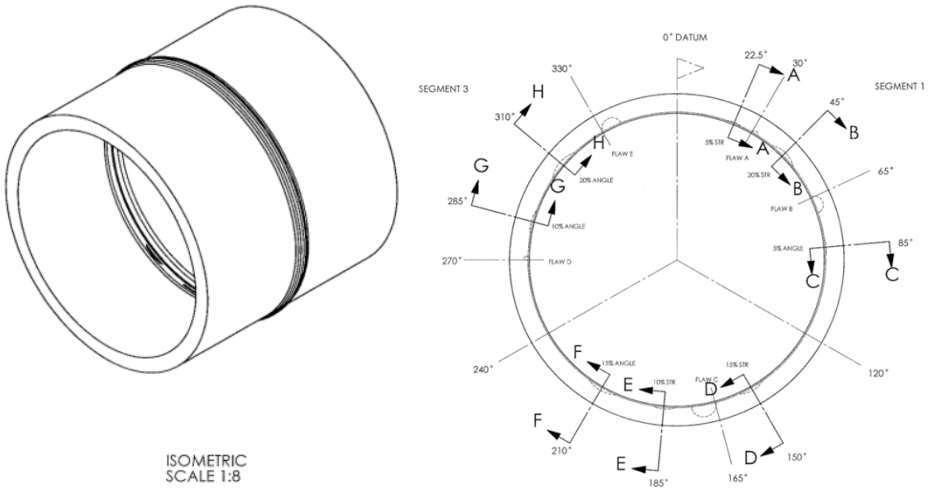
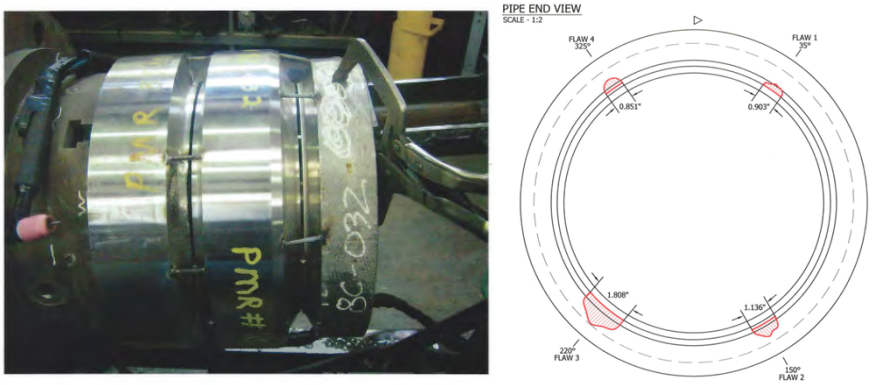


Figure A-3. Specimen 322-14-01P and flaw positions on the specimen 322-14-01P.



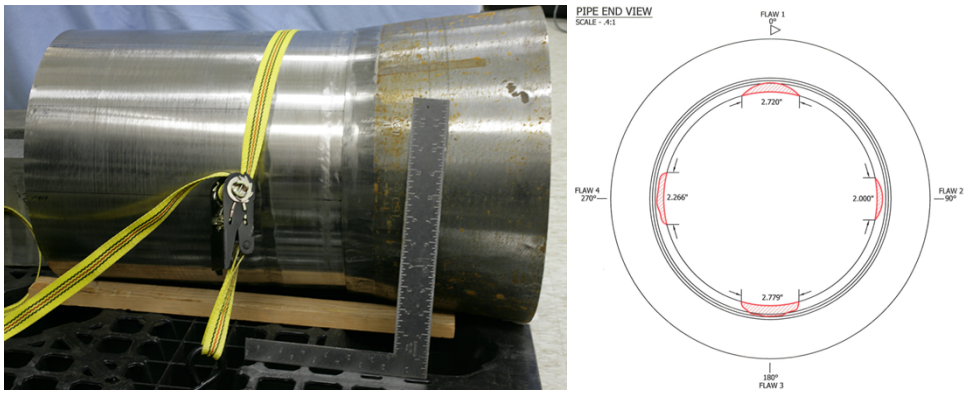
(a) **(b)**

Figure A-4. (a) Specimen 02-14-15 and (b) flaw positions on the specimen 02-14-15



(a) **(b)**

Figure A-5. (a) Specimen 8C-032 and (b) flaw positions on specimen 8C-032.



(a) **(b)**

Figure A-6. (a) Specimen 8C-091 and (b) flaw positions on specimen 8C-091

Table A-2. Summary of the flaw information

Specimen	Flaw	Type	OD to ID Thickness (mm)	Flaw length (mm)	Height (% thickness
19C-358-1	1	SC	73.2	101.7	30.1%
	2	SC	73.2	101.4	30.2%
	3	SC	73.2	101.6	30.2%
	4	SC	73.2	101.4	30.0%
19C-358-2	1	SC	28.6	100.6	29.2%
	2	SC	28.6	101.4	29.2%
	3	SC	28.6	101.4	29.4%
	4	SC	28.6	101.4	29.5%
322-14-01P	1	TFC	38.6	70.4	65.8%
	2	TFC	38.6	13.5	12.5%
	3	TFC	38.6	46.5	43.0%

02-24-15	A	TFC	36.0	10.7	15.0%
	B	TFC	35.3	30.5	43.0%
	C	TFC	35.7	43.6	64.0%
	a	SC	36.0	32.8	7.5%
	b	SC	35.9	65.2	28.4%
	d	SC	36.2	54.1	18.8%
	e	SC	35.8	43.7	12.0%
8C-032	1	TFC	38.2	22.9	20.0%
	2	TFC	36.0	28.9	40.0%
	3	TFC	38.2	45.9	60.0%
	4	TFC	36.0	21.6	30.0%
8C-091	1	EDM Notch	38.6	69.1	30.2%
	2	EDM Notch	39.9	50.8	17.6%
	3	TFC	38.8	70.6	36.4%
	4	TFC	39.7	57.6	23.2%

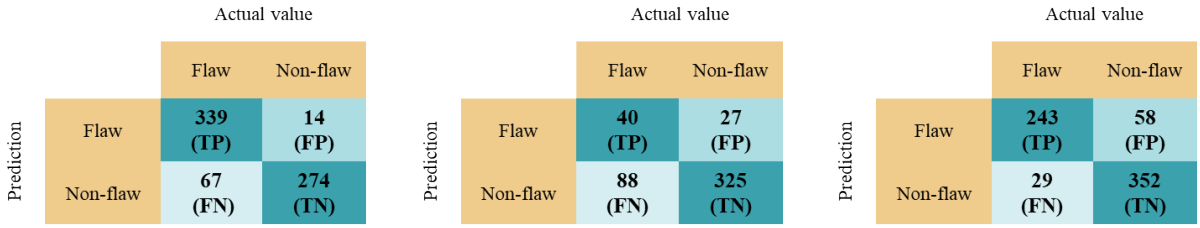
Table A-3. Cropping information for the SwRI 45° data (near side)

Specimen	19C-358-1	19C-358-2	322-14-01P	02-24-15	8C-032	8C-091
Original size	6734×381	4490×401	6286×217	4332×251	2628×189	3716×199
Area of interest	[500:5500] ×381	[500:2000] ×401	[500:2500] ×217	[500:2700] ×217	[500:2600] ×189	[500:3000] ×199
After downsampling	400×200	400×200	400×200	400×200	400×200	400×200

APPENDIX B. SUMMARY OF TEST RESULTS FROM MACHINE LEARNING

APPENDIX B. SUMMARY OF TEST RESULTS FROM MACHINE LEARNING

1. Using 19C-358-1 SwRI 45° (near side) as training data:



(a) 19C-358-2 (4 SC) (b) 322-14-01P (3 TFC) (c) 02-24-15 (3 TFC & 4 SC)

Figure B-1. Confusion matrices using the model trained by 19C-358-1.

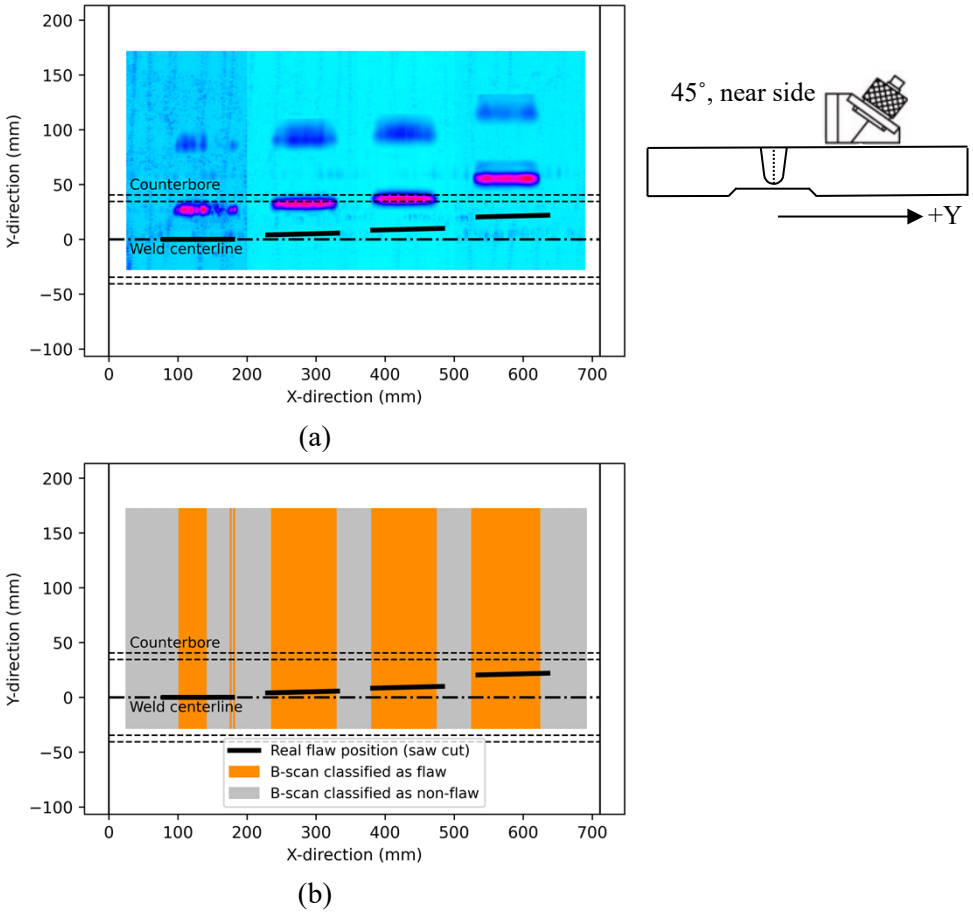
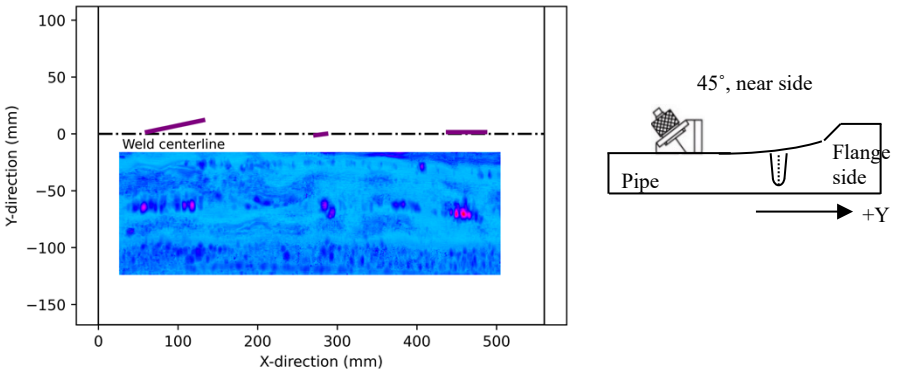
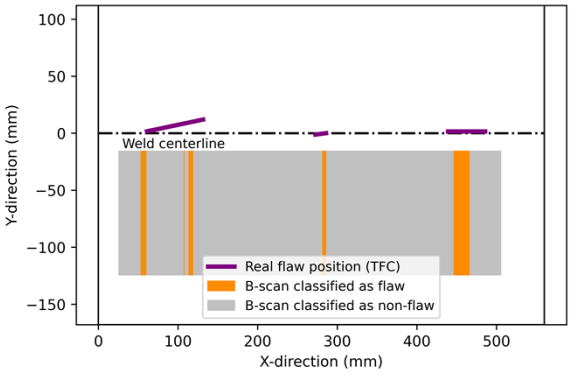


Figure B-2. Using 19C-358-1 SwRI 45° (near side) as training data: (a) Ultrasonic scanning image, (b) Test results on specimen 19C-358-2.

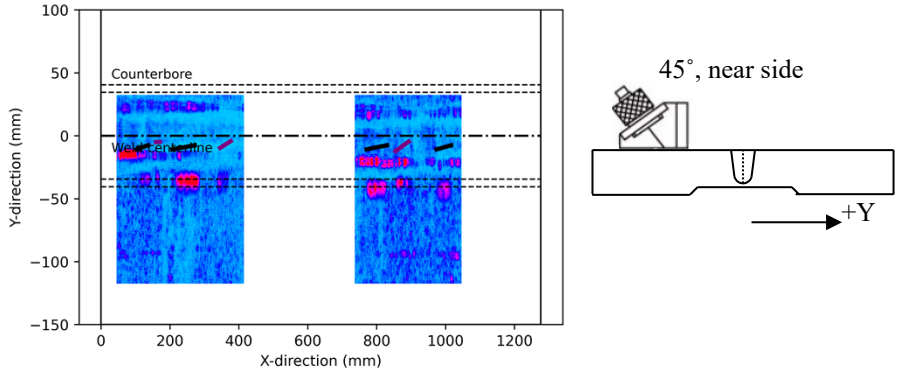


(a)

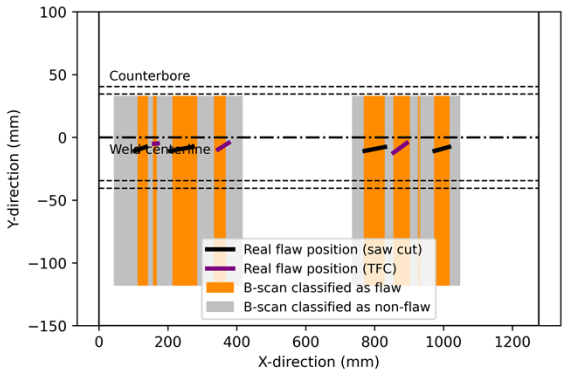


(b)

Figure B-3. Using 19C-358-1 SwRI 45˚ (near side) as training data: (a) Ultrasonic scanning image, (b) Test results on specimen 322-14-01P.



(a)



(b)

Figure B-4. Using 19C-358-1 SwRI 45˚ (near side) as training data: (a) Ultrasonic scanning image (b) Test results on specimen 02-24-15.

2. Using 19C-358-2 SwRI 45˚ (near side) as training data:

Prediction \ Actual value	Flaw	Non-flaw
Flaw	**241 (TP)**	**3 (FP)**
Non-flaw	**165 (FN)**	**285 (TN)**

(a) 19C-358-1 (4 SC)

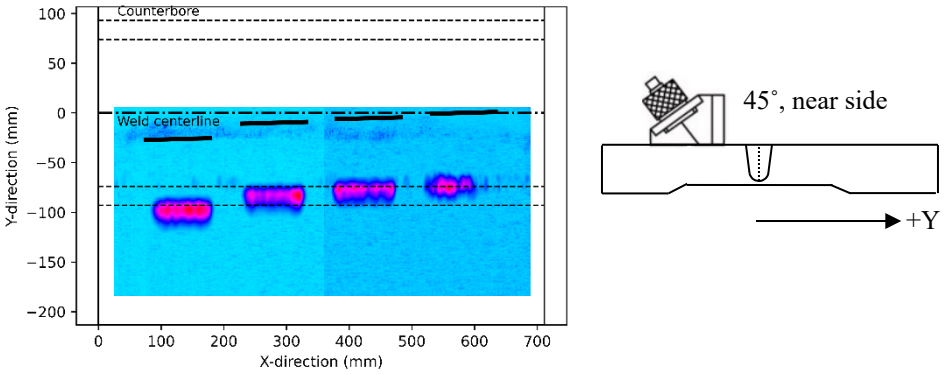
Prediction \ Actual value	Flaw	Non-flaw
Flaw	**99 (TP)**	**25 (FP)**
Non-flaw	**173 (FN)**	**385 (TN)**

(b) 322-14-01P (3 TFC)

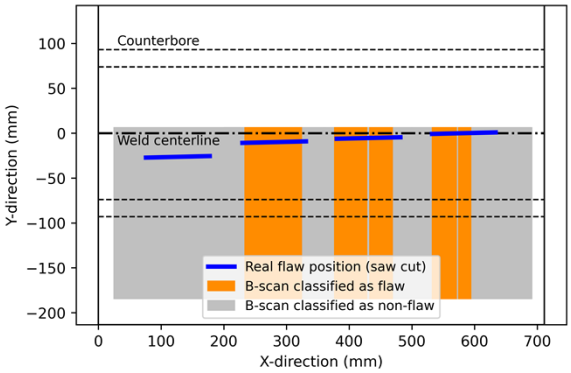
Prediction \ Actual value	Flaw	Non-flaw
Flaw	**7 (TP)**	**0 (FP)**
Non-flaw	**121 (FN)**	**352 (TN)**

(c) 02-24-15 (3 TFC & 4 SC)

Figure B-5. Confusion matrices using the model trained by 19C-358-2.

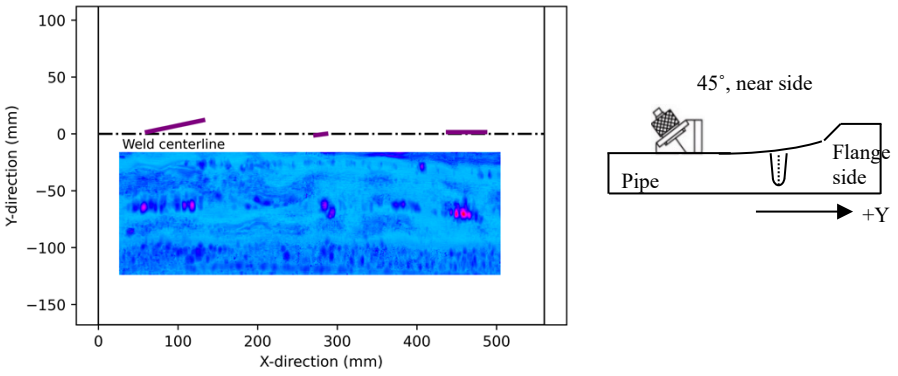


(a)

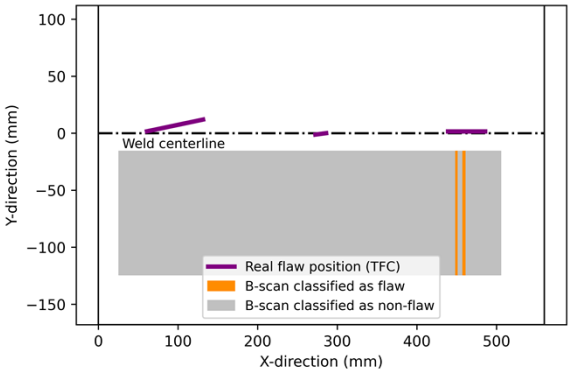


(b)

Figure B-6. Using 19C-358-2 SwRI 45˚ (near side) as training data: (a) Ultrasonic scanning image, (b) Test results on specimen 19C-358-1.

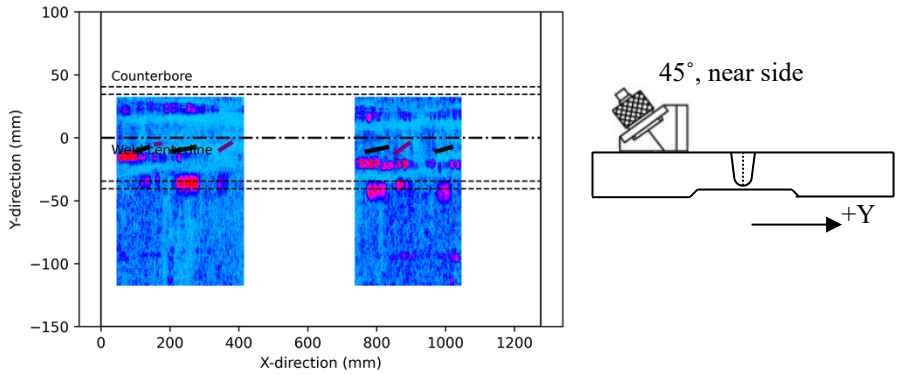


(a)

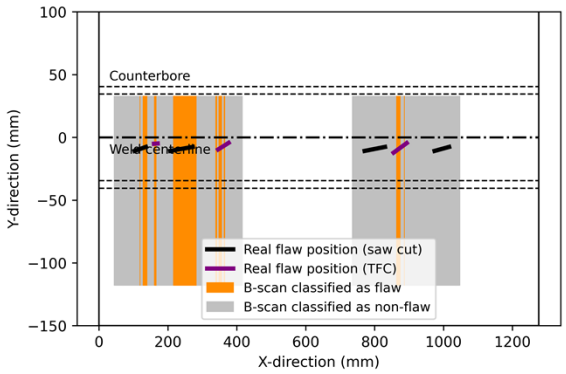


(b)

Figure B-7. Using 19C-358-2 SwRI 45˚ (near side) as training data: (a) Ultrasonic scanning image, (b) Test results on specimen 322-14-01P.



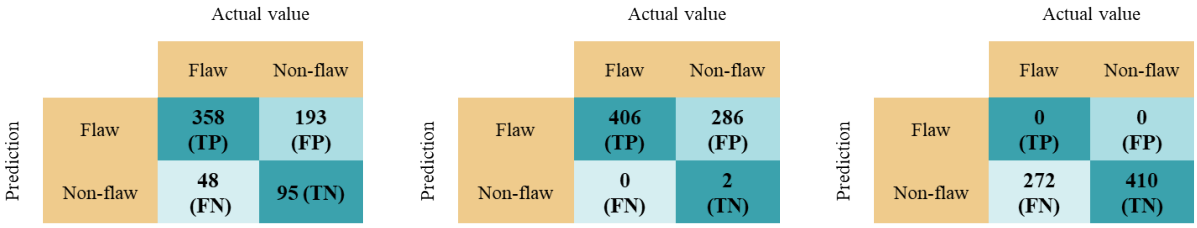
(a)



(b)

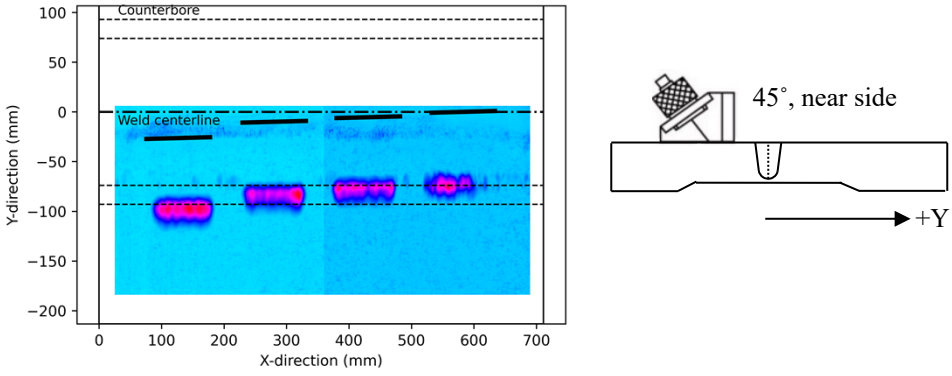
Figure B-8. Using 19C-358-2 SwRI 45˚ (near side) as training data: (a) Ultrasonic scanning image, (b) Test results on specimen 02-24-15.

3. Using 322-14-01P SwRI 45˚ (near side) as training data:

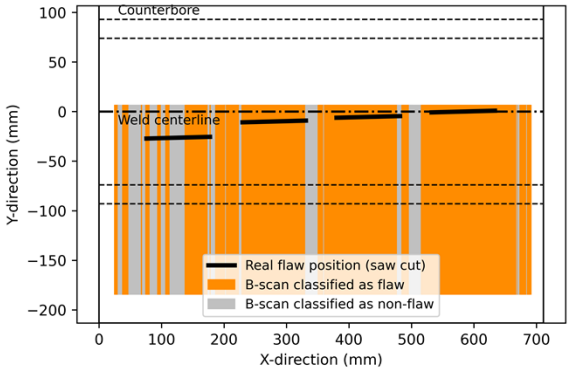


(a) 19C-358-1 (4 SC) (b) 19C-358-2 (4 SC) (c) 02-24-15 (3 TFC & 4 SC)

Figure B-9 Confusion matrices using the model trained by 322-14-01P (3 TFC).

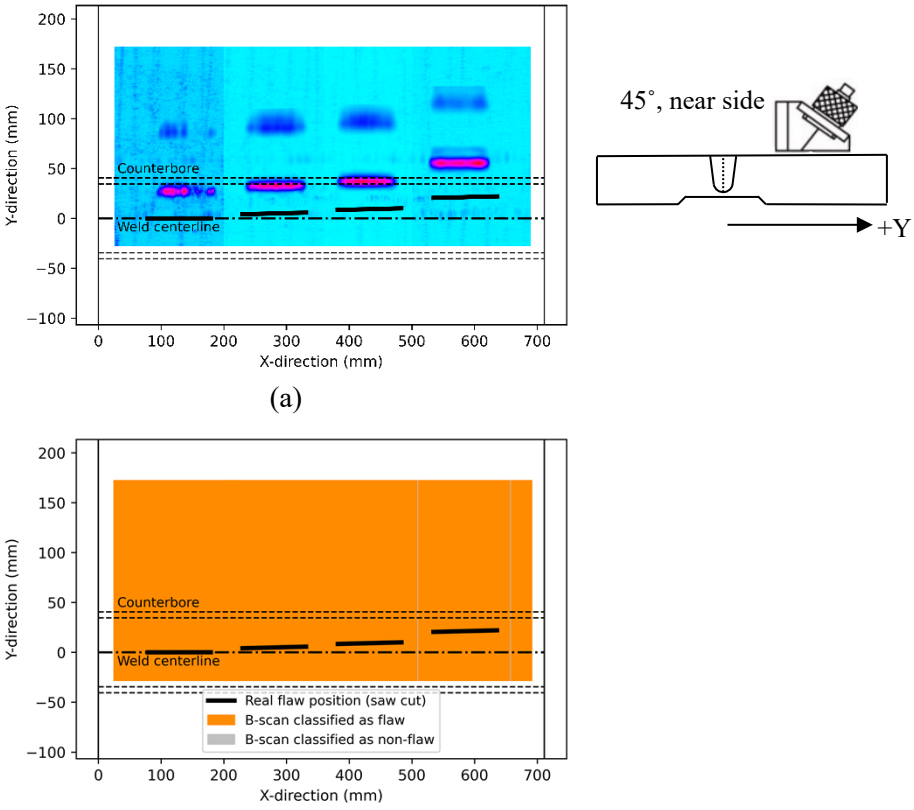


(a)



(b)

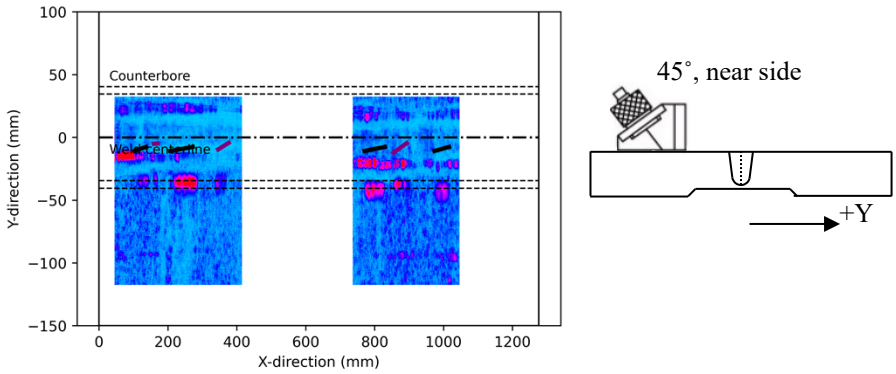
Figure B-10. Using 322-14-01P SwRI 45˚ (near side) as training data: (a) Ultrasonic scanning image, (b) Test results on specimen 19C-358-1.



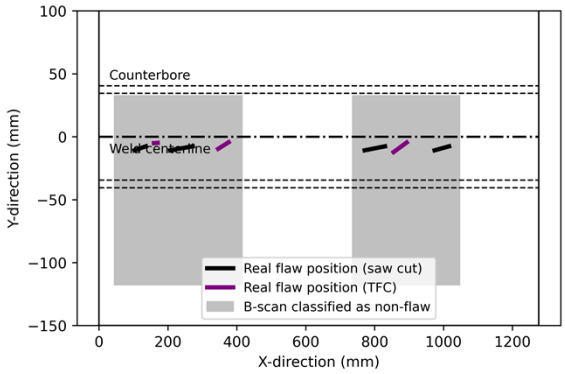
(a)

(b)

Figure B-11. Using 322-14-01P SwRI 45˚ (near side) as training data: (a) Ultrasonic scanning image, (b) Test results on specimen 19C-358-2.



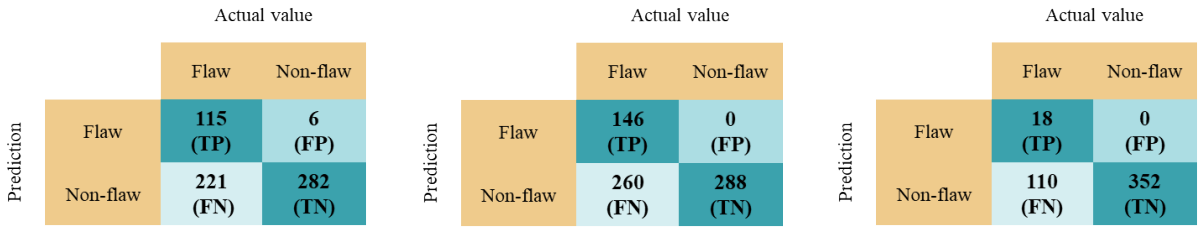
(a)



(b)

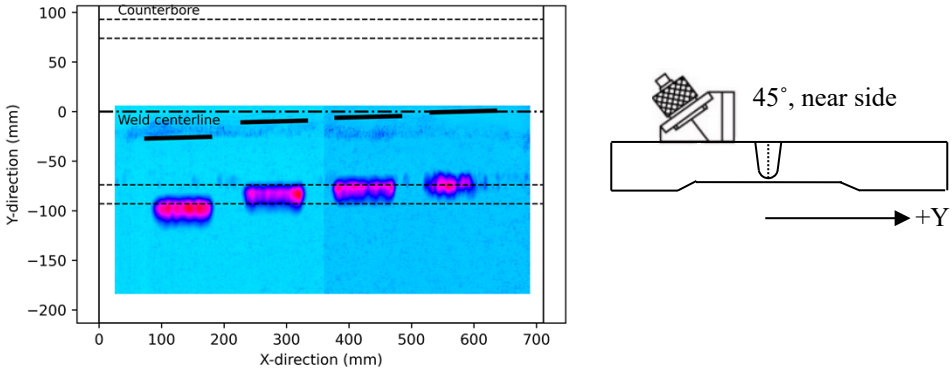
Figure B-12. Using 322-14-01P SwRI 45˚ (near side) as training data: (a) Ultrasonic scanning image, (b) Test results on specimen 02-24-15.

4. Using 02-24-15 (3TFC and 4SC) SwRI 45˚ (near side) as training data:

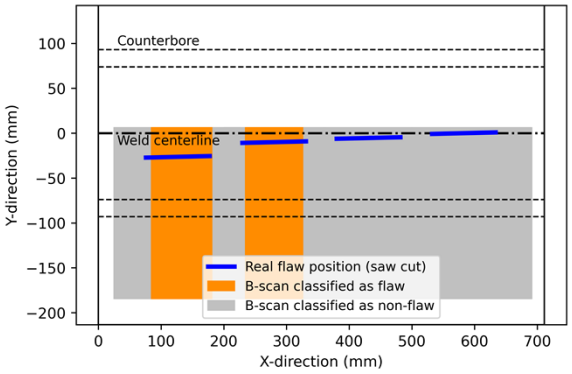


(a) 19C-358-1 (4 SC) (b) 19C-358-2 (4 SC) (c) 322-14-01P (3 TFC)

Figure B-13. Confusion matrices using the model trained by 02-24-15 (3 TFC and 4 SC).

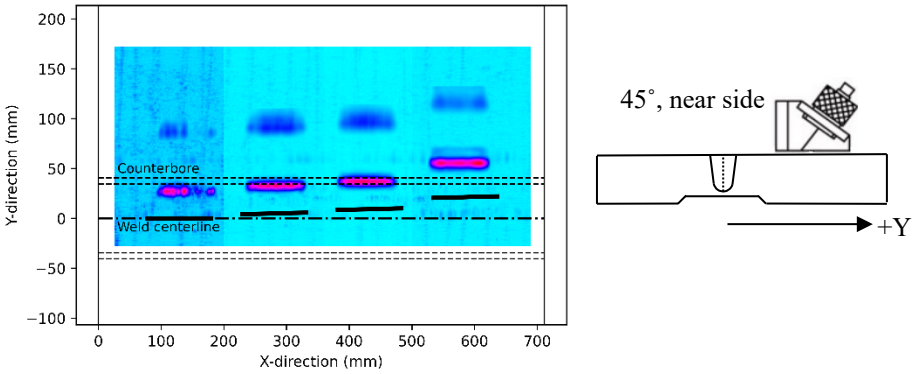


(a)

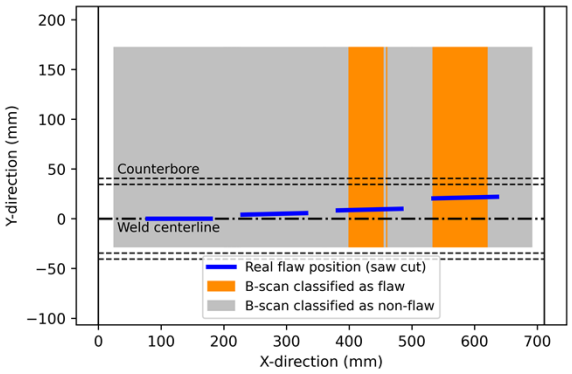


(b)

Figure B-14. Using 02-24-15 (3TFC and 4SC) SwRI 45˚ (near side) as training data: (a) Ultrasonic scanning image, (b) Test results on specimen 19C-358-1.

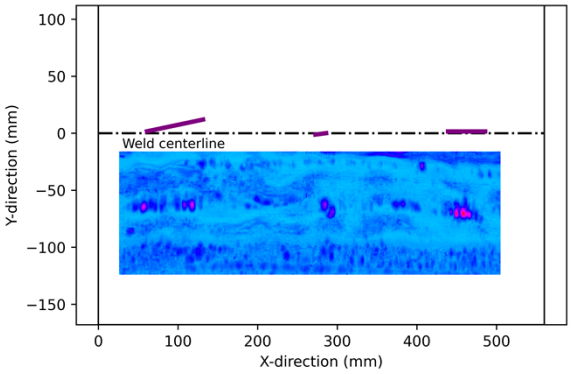


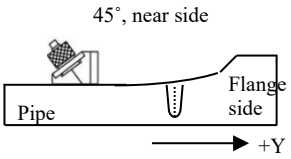
(a)



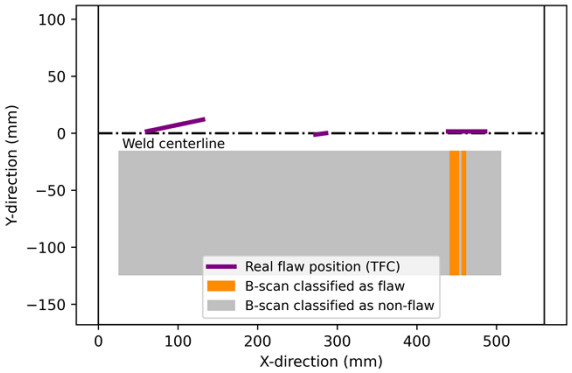
(b)

Figure B-15. Using 02-24-15 (3TFC and 4SC) SwRI 45˚ (near side) as training data: (a) Ultrasonic scanning image, (b) Test results on specimen 19C-358-2.





(a)



(b)

Figure B-16. Using 02-24-15 (3TFC and 4SC) SwRI 45˚ (near side) as training data: (a) Ultrasonic scanning image, (b) Test results on specimen 322-14-01P.

5. Using 02-24-15 (3TFC) SwRI 45° (near side) as training data:

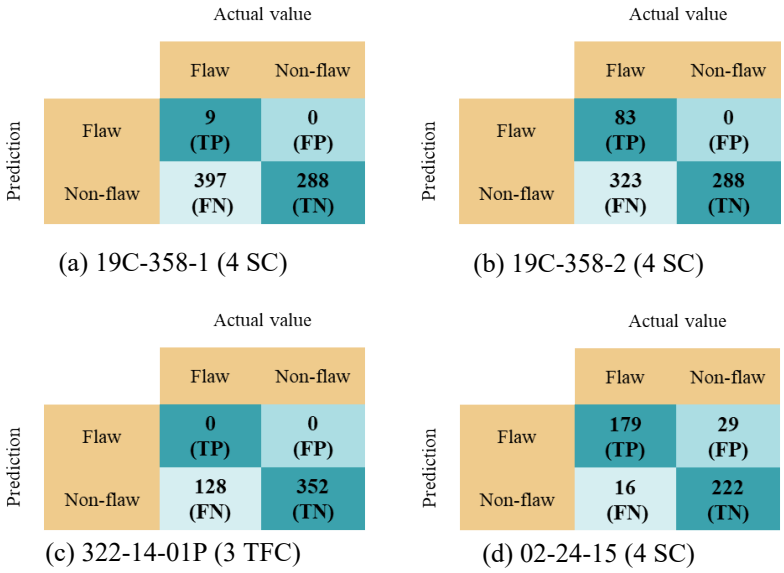
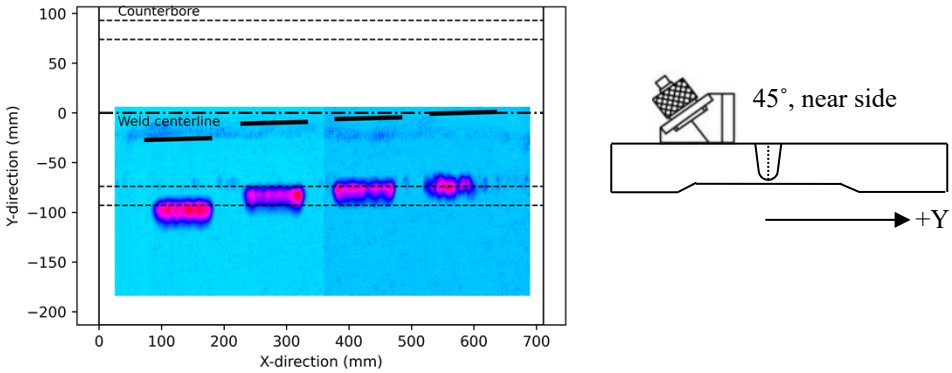
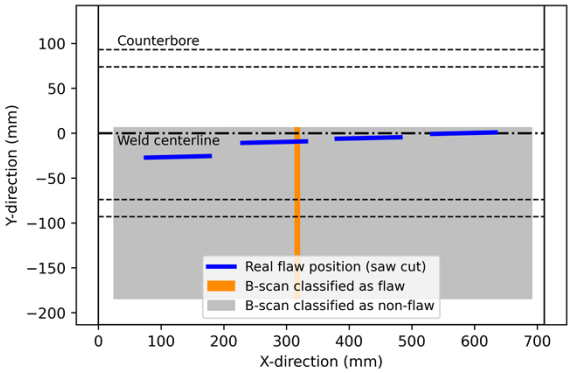


Figure B-17. Confusion matrices using the model trained by 02-24-15 (3 TFC).

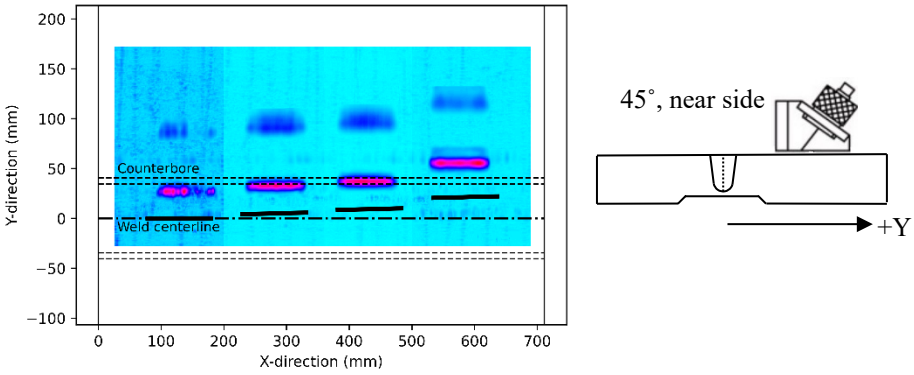


(a)

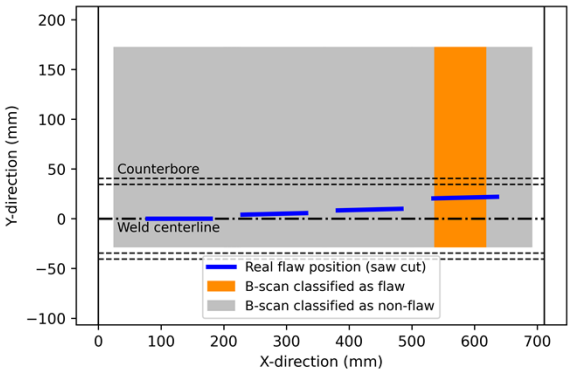


(b)

Figure B-18. Using 02-24-15 (3TFC) SwRI 45˚ (near side) as training data: (a) Ultrasonic scanning image, (b) Test results on specimen 19C-358-1.

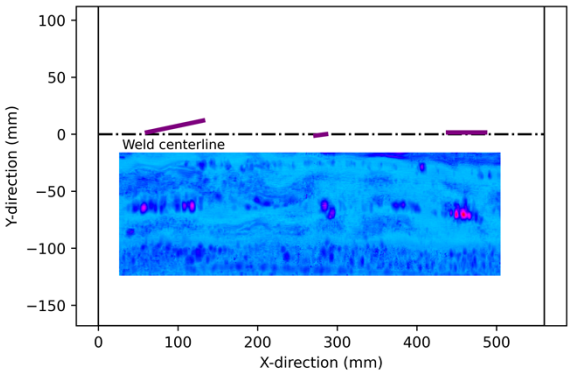


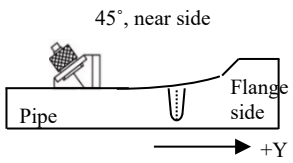
(a)



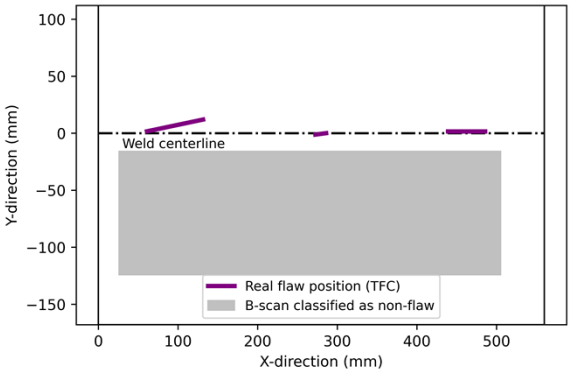
(b)

Figure B-19. Using 02-24-15 (3TFC) SwRI 45˚ (near side) as training data: (a) Ultrasonic scanning image, (b) Test results on specimen 19C-358-2.



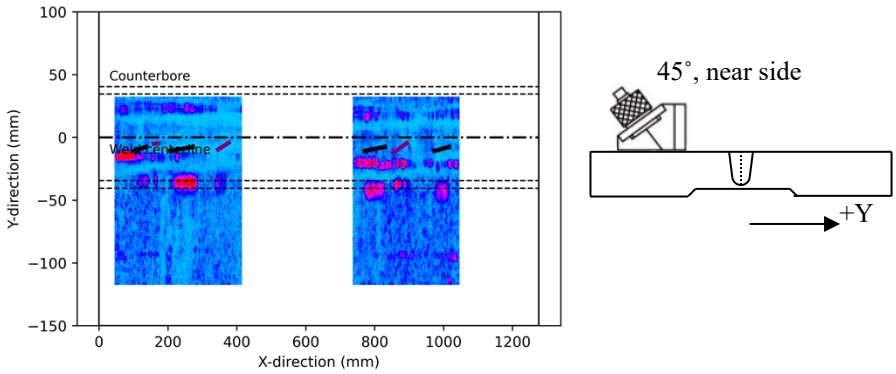


(a)

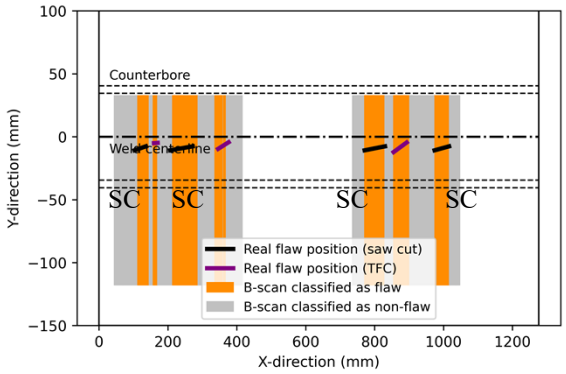


(b)

Figure B-20. Using 02-24-15 (3TFC) SwRI 45˚ (near side) as training data: (a) Ultrasonic scanning image, (b) Test results on specimen 322-14-01P.



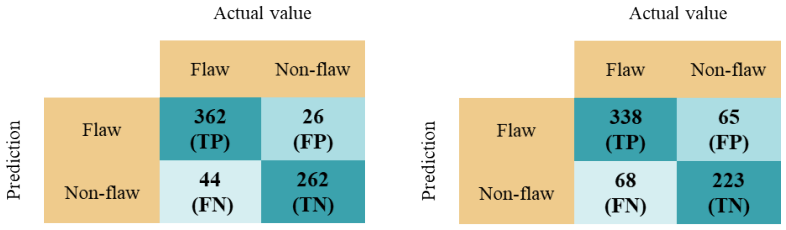
(a)



(b)

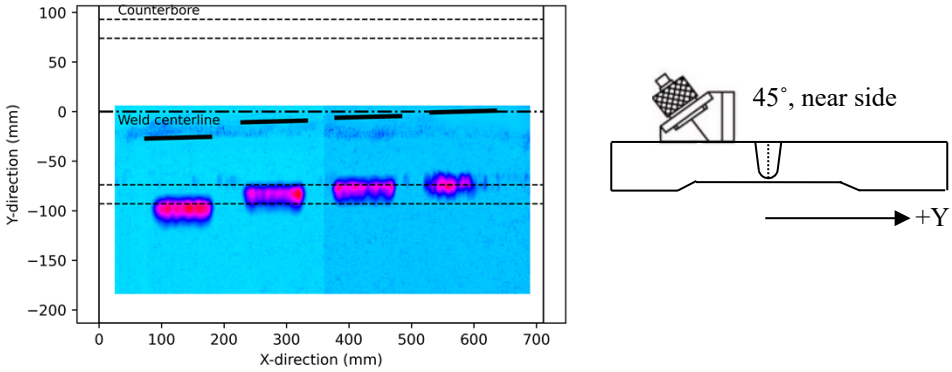
Figure B-21. Using 02-24-15 (3TFC) SwRI 45˚ (near side) as training data: (a) Ultrasonic scanning image, (b) Test results on specimen 02-24-15 (4 SC).

6. Using 322-14-01P (3TFC) SwRI 45˚ (initial training) and 02-24-15 (4SC) SwRI 45˚ (retraining) as training data:

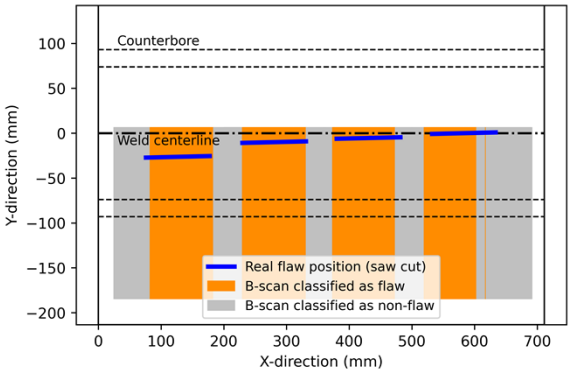


(a) 19C-358-1 (4 SC) (b) 19C-358-2 (4 SC)

Figure B-22. Confusion matrices using the model trained by 322-14-01P (3 TFC) first and then retrained by ss02-24-15 (4 SC).

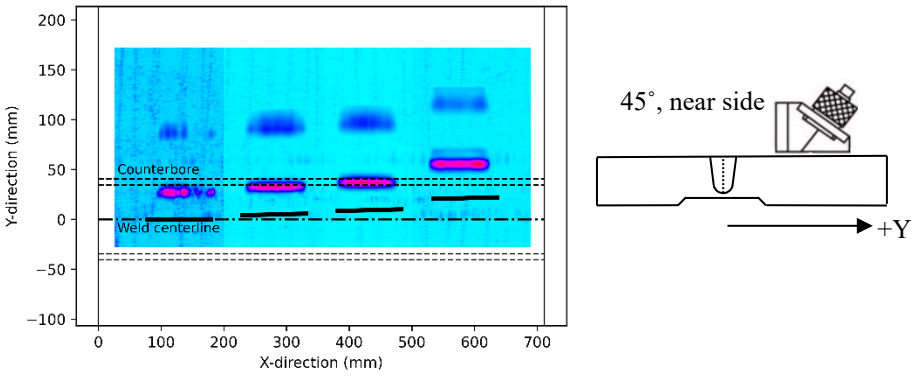


(a)

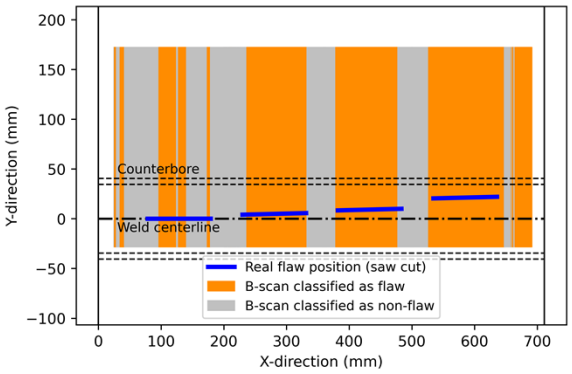


(b)

Figure B-23. Using 322-14-01P (3TFC) SwRI 45˚ (initial training) and 02-24-15 (4SC) SwRI 45˚ (retraining) as training data: (a) Ultrasonic scanning image, (b) Test results on specimen 19C-358-1.



(a)



(b)

Figure B-24. Using 322-14-01P (3TFC) SwRI 45° (initial training) and 02-24-15 (4SC) SwRI 45° (retraining) as training data: (a) Ultrasonic scanning image, (b) Test results on specimen 19C-358-2.