PNNL-SA-208872 Rev 1

Evaluation of Machine Learning Models for Automated Data Analysis in In-Service Nuclear Power Plant Inspections

Muthu Elen[1], Matt Prowant[1], Nick Conway[1], Joel Harrison[1], Aimee Holmes[1], Richard Jacob[1], Hongbin Sun[2], Pradeep Ramuhalli[2]

[1]Pacific Northwest National Laboratory (PNNL)
[2]Oak Ridge National Laboratory (ORNL)

This work was sponsored by the U.S. NRC
NRC COR: Isaac Anchondo-Lopez

EPRI NDE In Nuclear 2025
Los Angeles, CA
June 24th, 2025

Background and Scope

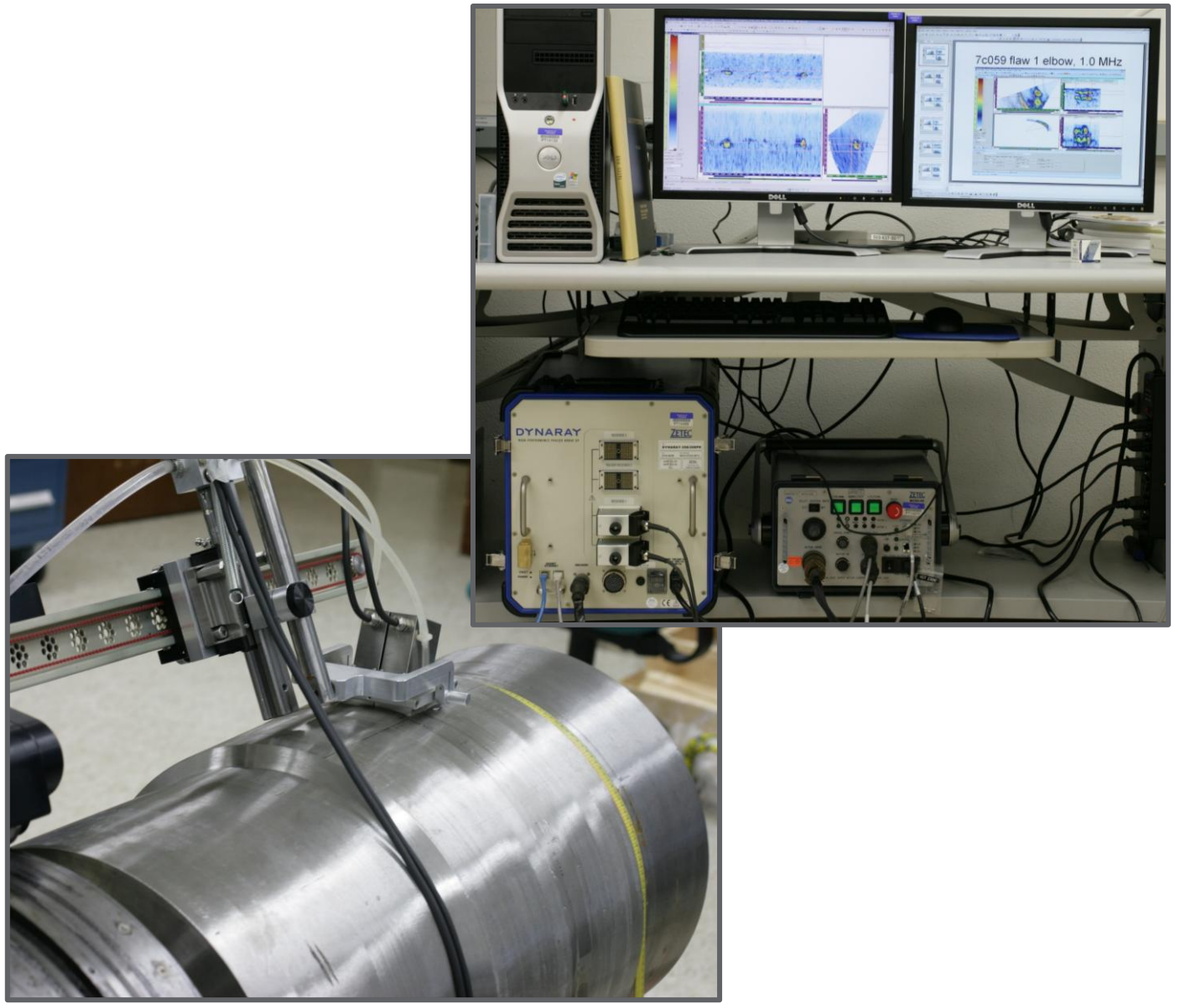
- The commercial nuclear power industry is facing a potential shortage of certified nondestructive evaluation (NDE) analysts to meet future in-service inspection demands.
- Automated data analysis (ADA) currently supports human inspectors in tasks such as eddy current evaluations for steam generator examinations.
- Machine learning (ML) systems are nearing the capability to pass performance demonstration tests for ultrasonic testing (UT) inspections of reactor pressure vessel upper head penetrations in nuclear power plants (NPPs).
- Current research and development is focused on assisted analysis (AA) of ADA versus fully automated examinations.
- This presentation will cover assessment of ML flaw detection on dissimilar metal weld (DMW) piping joints.

Commercial ML Systems for UT Inspections in NPPs

- Several commercial ML systems are being developed for assisted evaluation of NDE data in nuclear power plants.
- A few commercially available solutions (i.e., pre-trained) algorithms exist. One such system, referred to as ‘Box’ (made by TrueFlaw), is used in this research.
- PNNL’s assessments are aimed to identify factors affecting the use of assisted data analysis algorithms, including ML, for ASME Code-required ultrasonic in-service inspections.

PNNL Mockups and NDE Data Acquisition

- NDE inspections were performed on DMW mockups at PNNL.
- The majority of the mockups contained circumferentially oriented flaws consisting of thermal fatigue cracks (TFC) and a few electrical discharge machined (EDM) notches.
- Experimental data were collected according to the same procedure used to collect training data for the commercial ML system (Box).



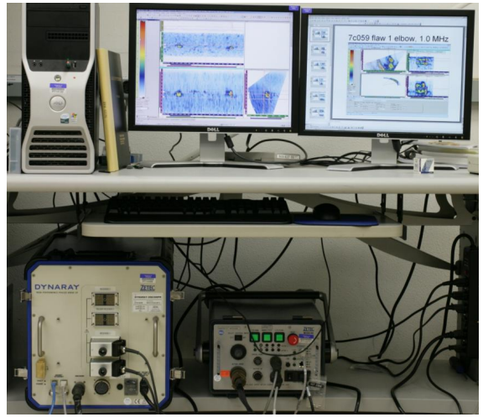
PNNL Mockups and NDE Data Acquisition

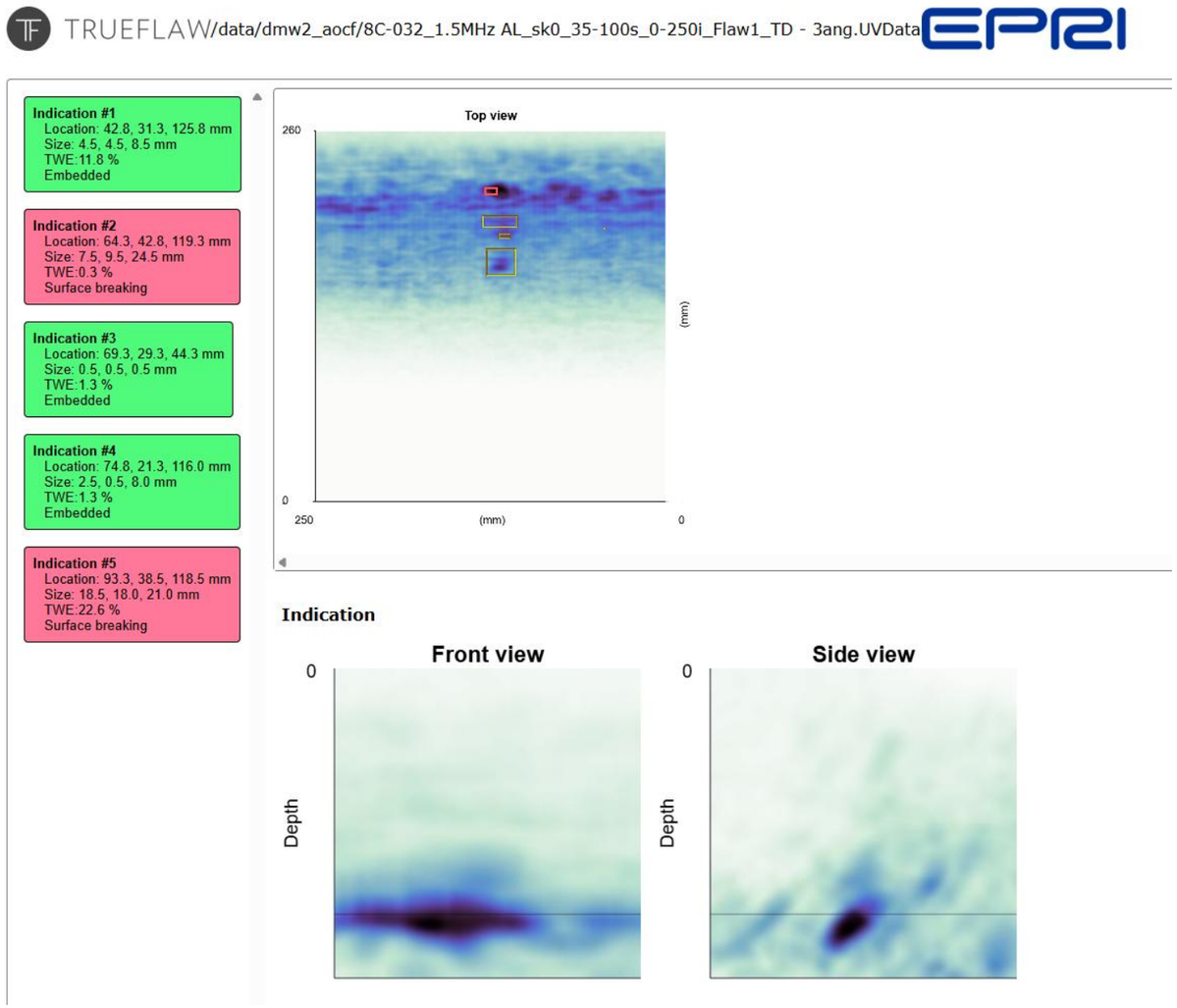
Specimen	Flaw	Flaw Type	Orientation	Flaw Length (in.)	Flaw Depth (in.)	%TW	Tilt	Thickness (in.)	Degree Location
8C-032	1	TFC	Circ	0.90	0.30	20%	0	1.51	35
	2	TFC	Circ	1.14	0.57	40%	19	1.42	150
	3	TFC	Circ	1.81	0.90	60%	0	1.51	220
	4	TFC	Circ	0.85	0.43	30%	13	1.42	325
8C-036	1	TFC	Circ	2.48	0.83	58%	30	1.42	47
	2	TFC	Circ	2.85	1.43	95%	0	1.50	135
	3	TFC	Circ	1.58	0.53	35%	0	1.50	195
	4	HIP'd EDM	Circ	1.58	0.53	35%	0	1.50	245
	5	TFC	Circ	2.26	1.13	75%	0	1.51	285
	6	HIP'd EDM	Circ	2.25	1.13	75%	0	1.50	330
9C-023	1	TFC	Circ	2.76	0.49	34%	2	1.45	0
	2	TFC	Circ	2.01	0.26	19%	8	1.41	90
	3	TFC	Circ	2.76	0.33	24%	4	1.39	180
	4	TFC	Circ	2.26	0.16	11%	12	1.43	270
9C-034	1	TFC	Circ	2.02	0.18	9%	0	2.09	0
	2	TFC	Circ	2.03	0.42	20%	0	2.09	90
	3	TFC	Circ	2.52	0.26	12%	0	2.09	180
	4	TFC	Circ	2.51	0.36	17%	0	2.09	270
10C-011	1	TFC	Circ	1.50	0.65	34%	3	1.89	20
	2	TFC	Circ	1.71	0.78	41%	2	1.89	90
	3	TFC	Circ	2.01	0.91	48%	3	1.89	160
	4	TFC	Circ	2.21	1.04	55%	3	1.89	230
	5	TFC	Circ	2.51	1.17	62%	3	1.89	310
14C-146	4	TFC	Circ	2.965	0.522	15.7%	0	3.323	

	sk0	sk180
10C-011	CS	CASS
8C-032	CS	safe (WSS)
8C-036	CS	safe (WSS)
9C-023	CS	WSS
9C-034	CS	CASS
14C-146	CASS	CS

- DMW specimens selected were 12-14 in. OD pipe to nozzle specimens, with the addition of one 36 in. OD specimen (14C-146)
  - Two specimens also contained full structural weld overlays (9C-034, 10C-011)
- Data were collected from both sides of the target welds and processed through Box
  - Scans from CASS were excluded in the analysis results due to no applicable procedure

ML Model Evaluation





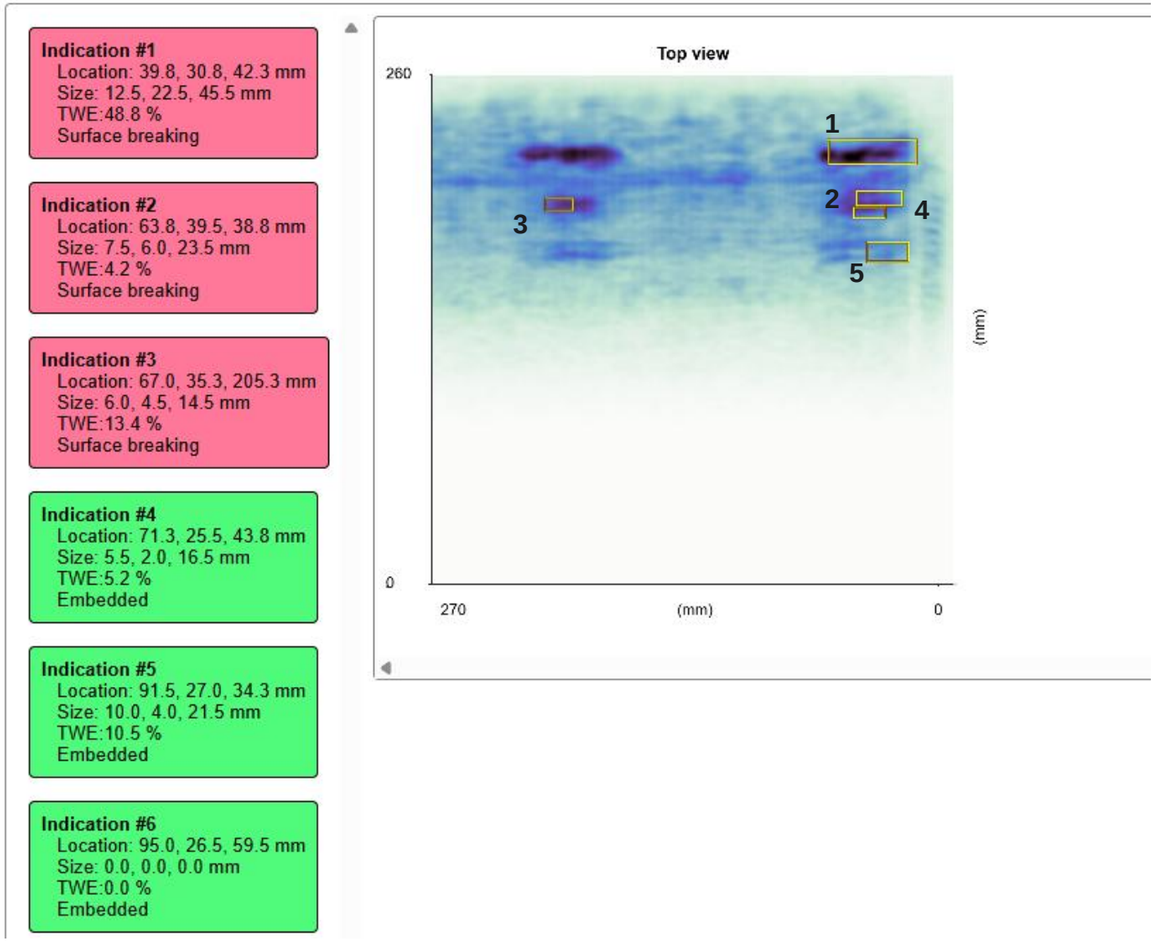
- The scan data files were input to the latest DMW model in Box, received in April 2025, and the output HTML files, consisting of flaw indications were collected.
- Each indication in the output files were validated with the true-state information by UT Level III Inspector, as per grading criteria.

Performance Evaluation of Box

- Grading criteria were developed for the evaluation of the data. Each “indication" can be labelled as
  - Detection (true positive) - if the indication correctly identified the flaw
  - False call (false positive) - if the indication identified a non-flaw as flaw
  - Reject - (1) If indication identified the same flaw multiple times, as only one indication can be considered as true call; (2) if size is 0 in all directions
  - Missed detection (false negatives) - a flaw was not identified as indication
  - Due to certain limitations with the location information in indications, the bounding boxes with tolerances were not developed, and hence an Intersection of Union (IoU) metric could not be obtained
  - True negatives – an unflawed region identified as non-flaw is a commonly used metric in ML; however, it is not applicable to detection-based algorithms
  - Additionally, it was noted that few false calls are of size 0-2 mm, but a few of these false calls appeared to be an indication of flaw. Hence, these small indications were not rejected and were considered as false calls

Example of Output File

TRUEFLAW/data/dmw2_aocf/8C-036_1.5MHz AL_sk0_35-100s_550-820i_Flaw3-4_TD



Flaw 3 – TFC
Flaw 4 – Hip'd EDM

- Indication #1 - detection of flaw 3
- Indication #2 - part of flaw 3
- Indication #3 - detection of flaw 4
- Indication #4 - part of flaw 3
- Indication #5 - part of flaw 3
- Indication #6 – size of 0.0,0.0,0.0

As per PNNL's Grading criteria,

Detection – Indications #1, #3
Reject - #2, #4, #5 and #6

Performance Evaluation of Box

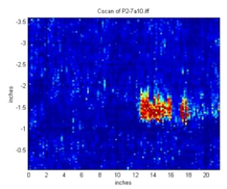
- Based on the evaluation, the following metrics were calculated:
  - Detection rate—the total number of correctly identified flaws divided by the total number of flaws
  - Probability of detection (POD)—the likelihood that flaws will be correctly identified within the inspection
  - False call probability (FCP)—the likelihood that flaws will be incorrectly identified within the inspection
  - Missed detections—the number of flaws that were not identified within the inspection
  - Other ML metrics such as Accuracy, Precision, Recall and F1-Score
- Results are as follows

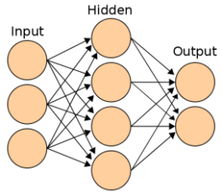
Metrics	Value	
Detection rate	92.5%	
95% Bounds	81.46%	99.19%
FCP	44%	
Missed detections	3	

Metrics	Value
Accuracy	0.42
Precision	0.44
Recall	0.93
F1-Score	0.60

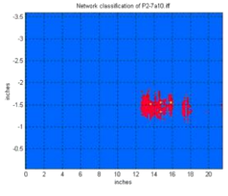
* The new model received earlier in June 2025 has improved with100% detection rate, identifying all the flaws. The # of false calls was higher.

Overview of Other ML Techniques

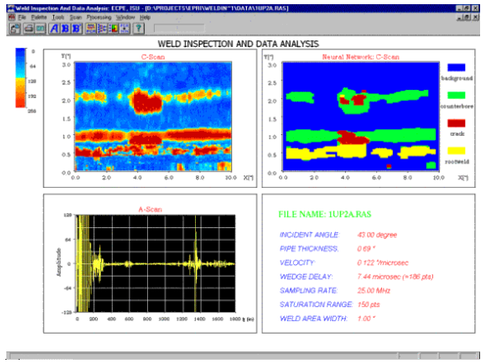


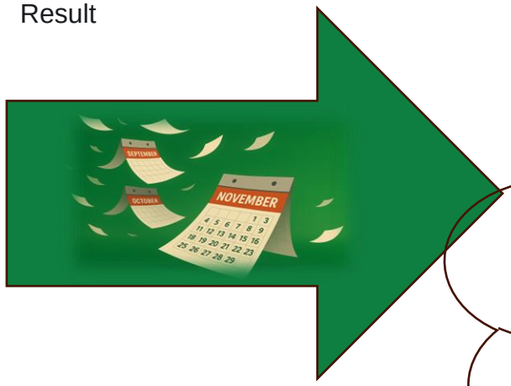


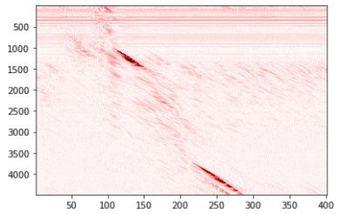
Single-element UT NDE Data

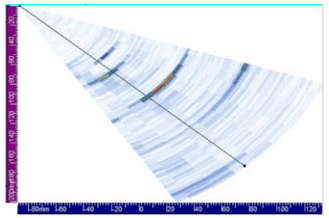


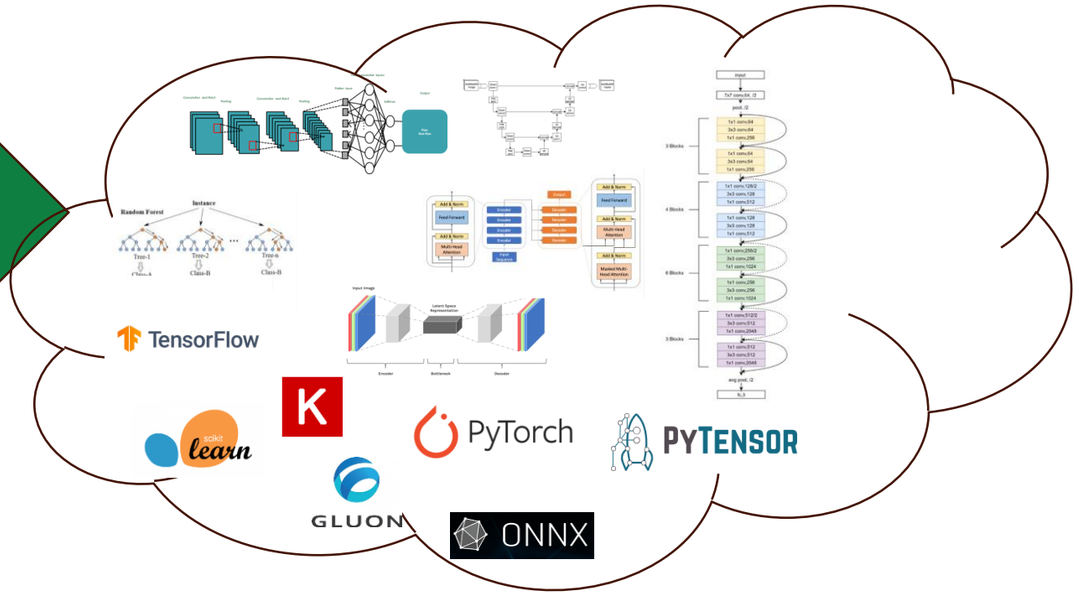
NN Classification Result











Examples of Machine Learning Applied to IGSCC (Examples from open specimens)*

*Spanner et al, 2nd Int'l. Conf. NDE in Relation to Structural Integrity for Nuclear and Pressurized Components, New Orleans May 2000

ML for UTNDE Data Analysis – circa 2000

UTNDE and ML Landscape – circa 2025

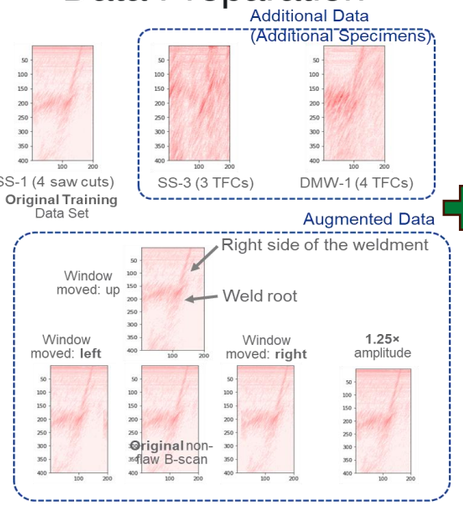
Disclosure: This slide contains some AI-generated graphics.

Evaluation Approach and Data

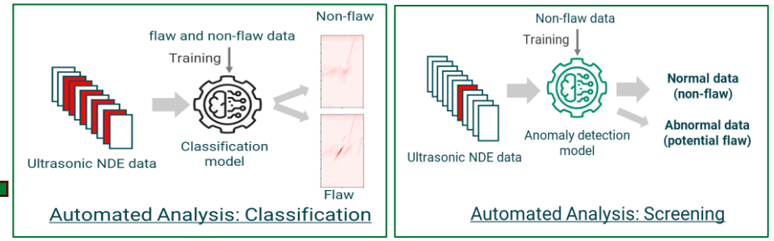
Specimen	Description	Flaw	Type	Flaw length (mm)	Height (% thickness
A(19C-358-1)	SS Plate	1	Saw cut	101.7	30.1%
		2	Saw cut	101.4	30.2%
		3	Saw cut	101.6	30.2%
		4	Saw cut	101.4	30.0%
B(19C-358-2)	SS Plate	1	Saw cut	100.6	29.2%
		2	Saw cut	101.4	29.2%
		3	Saw cut	101.4	29.4%
		4	Saw cut	101.4	29.5%
C(322-14-01P)	SS pipe section	1	TFC	70.4	65.8%
		2	TFC	13.5	12.5%
		3	TFC	46.5	43.0%
D(02-24-15)	SS pipe section	A	TFC	10.7	15.0%
		B	TFC	30.5	43.0%
		C	TFC	43.6	64.0%
		a	Saw cut	32.8	7.5%
		b	Saw cut	65.2	28.4%
		d	Saw cut	54.1	18.8%
		e	Saw cut	43.7	12.0%
E(8C-032)	DMW pipe	1	TFC	22.9	20.0%
		2	TFC	28.9	40.0%
		3	TFC	45.9	60.0%
		4	TFC	21.6	30.0%
F(8C-091)	DMW pipe	1	EDM notch	69.1	30.2%
		2	EDM notch	50.8	17.6%
		3	TFC	70.6	36.4%
		4	TFC	57.6	23.2%
G (21C-303-1)	SS plate	1	EDM notch	50.8	15.0%
		2	EDM notch	75.9	29.6%
		3	TFC	49.8	14.8%
		4	TFC	75.7	26.3%
H (21C-303-3)	SS plate	1	EDM notch	50.8	14.3%
		2	EDM notch	75.2	30.3%
		3	TFC	51.8	16.0%
		4	TFC	77.0	29.3%

Flaws in Reference Dataset (34 total: 12 saw cuts, 16 thermal fatigue cracks, 6 EDM notches)

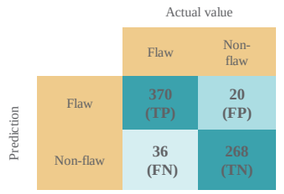
Data Preparation



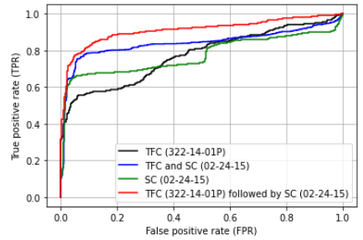
Problem Formulation



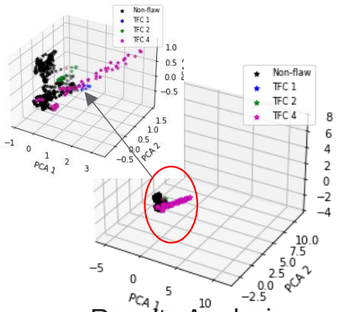
Model Selection, Parameter Variability Studies



Confusion matrix



ROC Curves

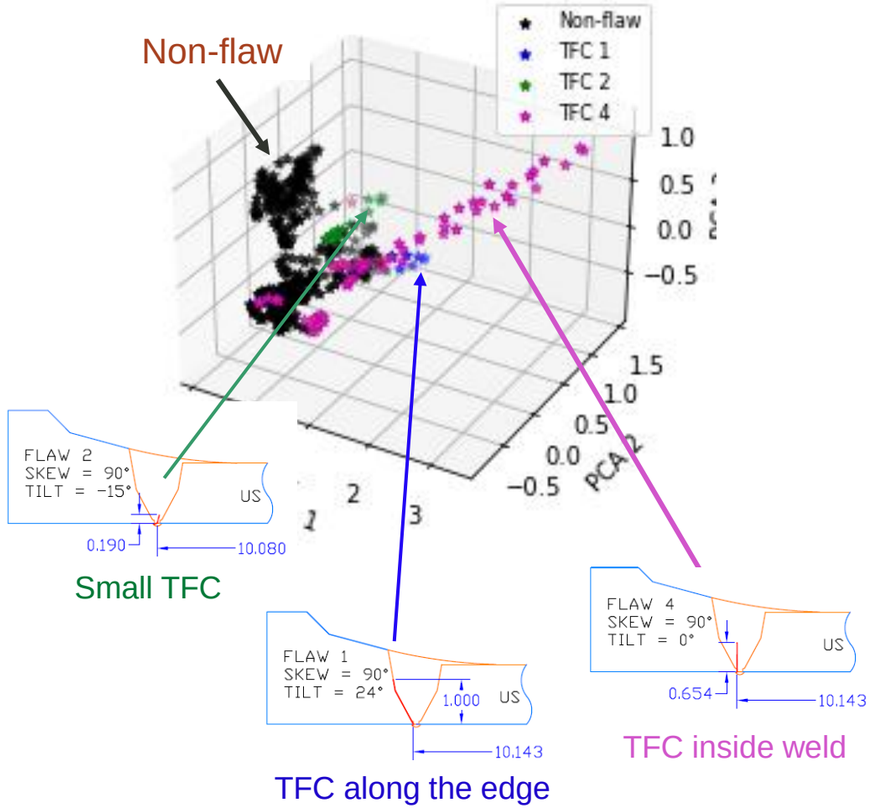


Results Analysis

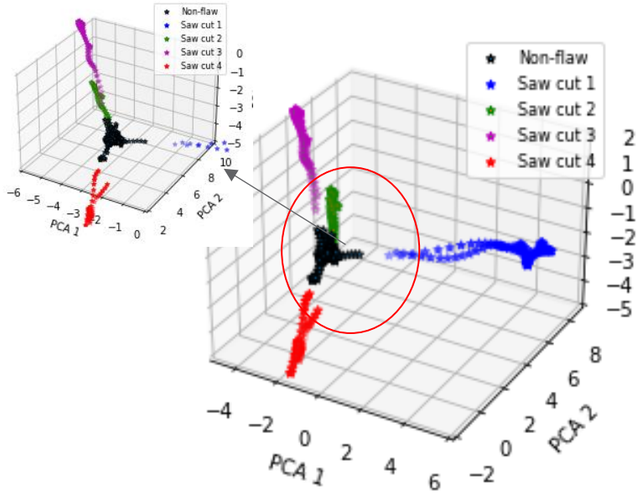
ML Results

Review of Results to Date and Findings

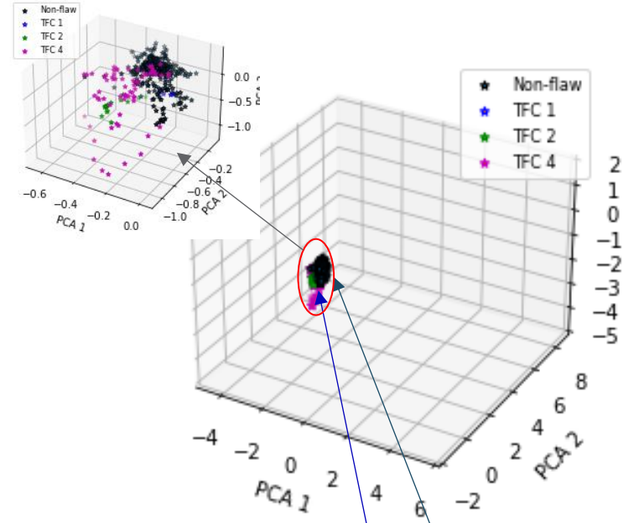
Principal Component Analysis (PCA) based on B-scan images (using three components)



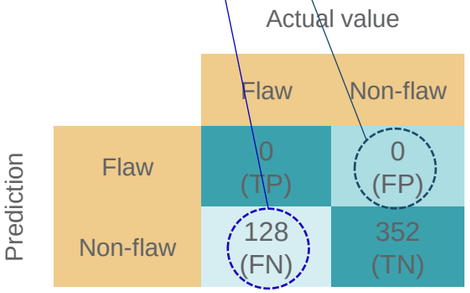
PCA, Specimen C (SwRI 45˚)



PCA, Specimen A (PAUT 45˚), Used for Model Training

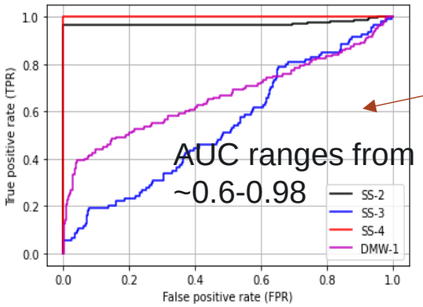


PCA Transformation from Specimen A, applied to Specimen C (PAUT 45˚)

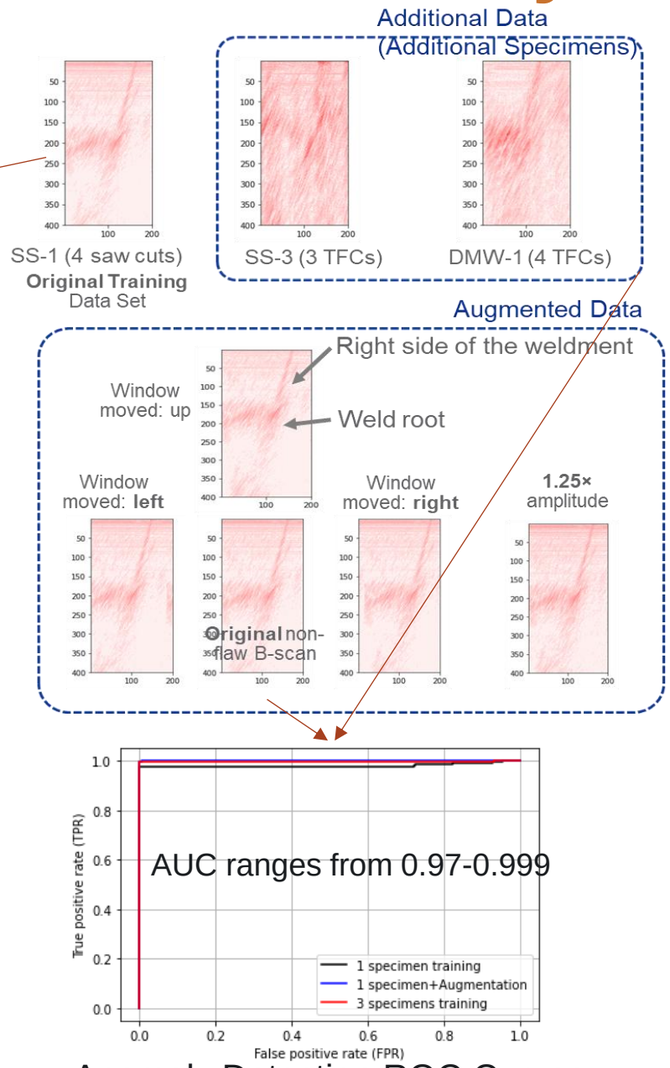


Accuracy=0.73,
TPR=0, FPR=0
ML results for Specimen C (CNN model trained with Specimen A data)

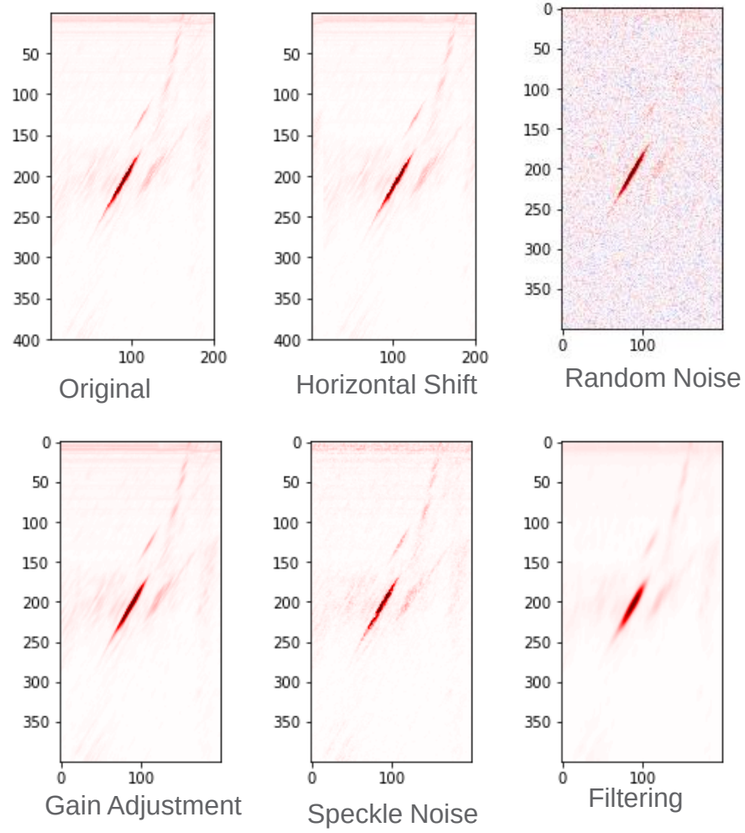
Richness of Training Data is Important for ML Accuracy



Anomaly Detection ROC Curves: before data augmentation



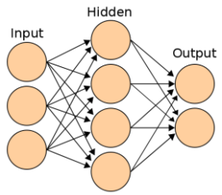
Anomaly Detection ROC Curves: after data augmentation



Other Impacts of Data Augmentation Being Evaluated

Performance for Different ML models

Supervised learning
CNN classification model



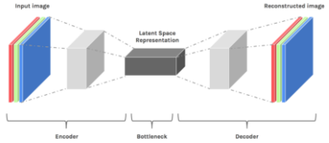
Training: specimen A
Testing: specimen D

Actual value

Prediction	Flaw	Non-flaw
Flaw	262 (TP)	0 (FP)
Non-flaw	8 (FN)	327 (TN)

TFR=0.97, FPR=0

Supervised learning
CNN Anomaly Detection model



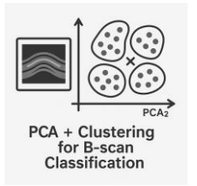
Training: specimen A
Testing: specimen D

Actual value

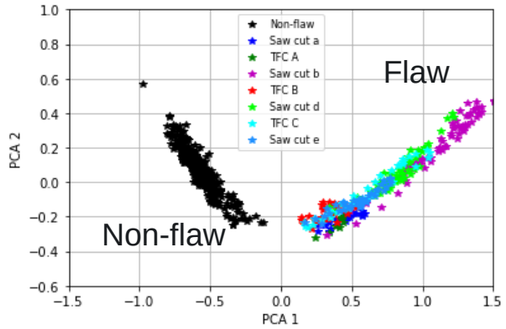
Prediction	Flaw	Non-flaw
Flaw	**268 (TP)**	**183 (FP)**
Non-flaw	**1 (FN)**	**144 (TN)**

TFR=0.99, FPR=0.56

Unsupervised learning
(feature extraction+ PCA + clustering)



specimen D



TFR=1, FPR=0

Summary

- Machine learning models were evaluated with ultrasonic NDE data collected on weld mockups (austenitic and dissimilar metal welds).
- The initial results suggest that ML has the potential for supporting automated NDE data analysis if applied appropriately.
- Results indicate several factors, including data richness, play a role in ML accuracy.
- The model also gives multiple indications of a flaw, which can include mode converted responses and tips signals.
- Limitations include having a high false call rate.

Future Work

- ML
  - Investigating data verification and validation methods.
  - Evaluating data augmentation approaches and simulation data sets for training ML.
  - Examining model explainability approaches.
- Evaluation
  - Evaluating data with new models of box.
  - Other statistical metrics for ML performance, including POD.
- Use cases
  - Retraining ML DMW models to assess need for site-specific requirements.
  - Reactor pressure vessel upper head penetrations.

Pacific Northwest
NATIONAL LABORATORY

Thank you