Oak Ridge National Laboratory

Pacific Northwest National Laboratory

Evaluation of Machine Learning (ML) for Automated Ultrasonic NDE Data Analysis

Hongbin Sun, Noel Nelson, Pradeep Ramuhalli

2026 Industry/NRC NDE Technical Information Exchange Meeting
January 21-22, 2026

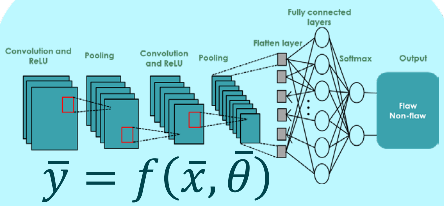
This work was sponsored by the U.S. NRC
Chinthaka Silva, NRC COR

U.S. DEPARTMENT OF ENERGY

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

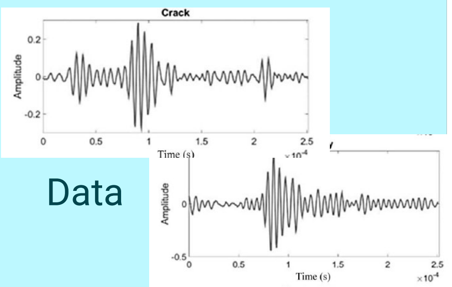
Pub. ID: 251268

Objective: Provide technical basis to support regulatory decisions and Code actions on automated data analysis for NDE



$$\bar{y} = f(\bar{x}, \bar{\theta})$$

Model and Model Parameters

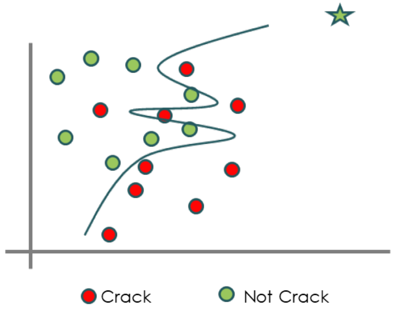


Data

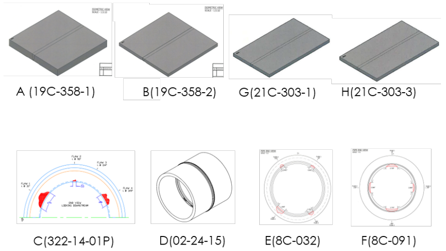
https://www.zetec.com/blog/destructive-and-nondestructive-testing-of-welds-how-ndt-ensures-quality/

Automated Analysis/AI/ML Purpose: Application

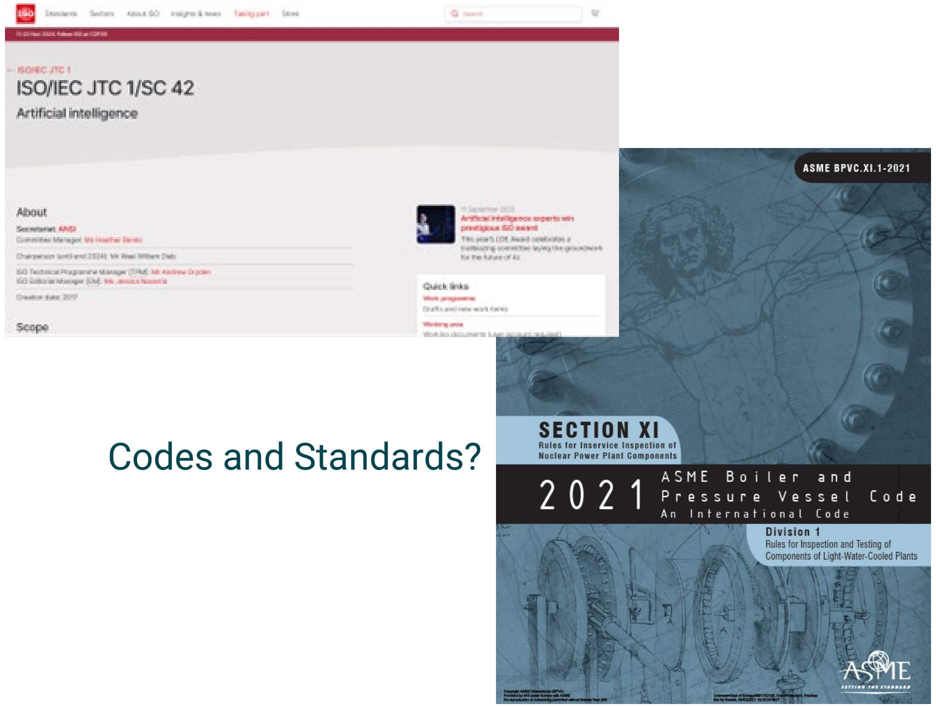
<u>Drivers Influencing ML Performance</u>



Validation and Qualification Requirements?



Data Requirements?



← ISO/IEC JTC 1

ISO/IEC JTC 1/SC 42

Artificial intelligence

About

Secretariat: ANSI

Committee Manager: [illegible]

Chairperson [illegible]

ISO Technical Programme Manager (TPM): [illegible]

ISO Editorial Manager (EM): [illegible]

Creation date: 2017

Scope

Artificial intelligence experts win prestigious ISO award

This year's LDE Award celebrates a trailblazing committee laying the groundwork for the future of AI

Quick links

Work programme

Drafts and new work items

Working area

Working documents (user account required)

ASME BPVC.XI.1-2021

SECTION XI

Rules for Inservice Inspection of Nuclear Power Plant Components

2021 ASME Boiler and Pressure Vessel Code

An International Code

Division 1

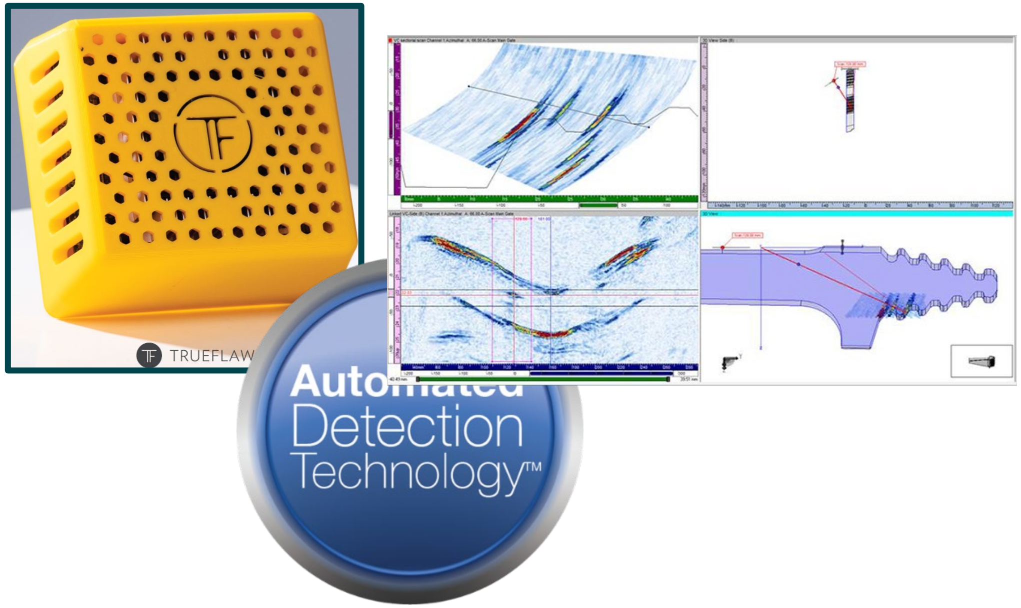
Rules for Inspection and Testing of Components of Light-Water-Cooled Plants

ASME

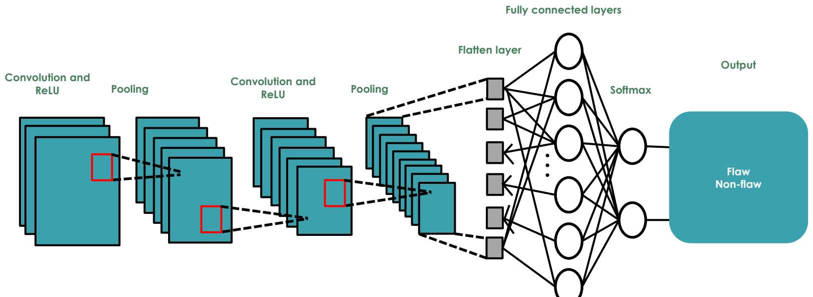
SETTING THE STANDARD

Codes and Standards?

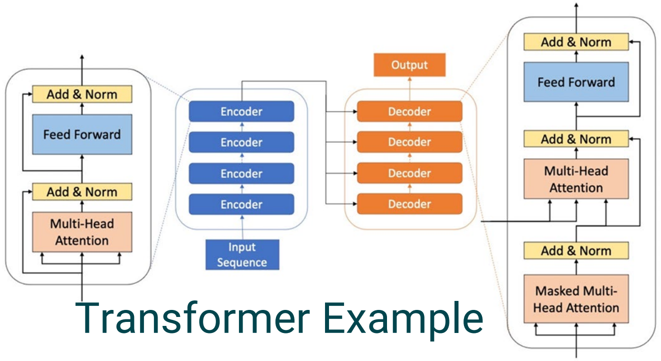
Research is focused on capability and limitation evaluations of commercial and research-grade machine learning methods



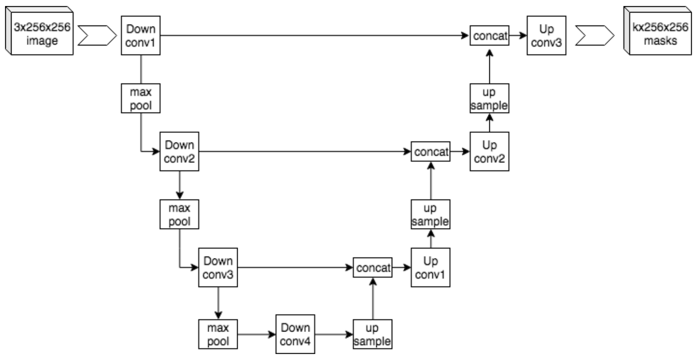
Examples of Commercial Automated Analysis/Machine Learning Tools



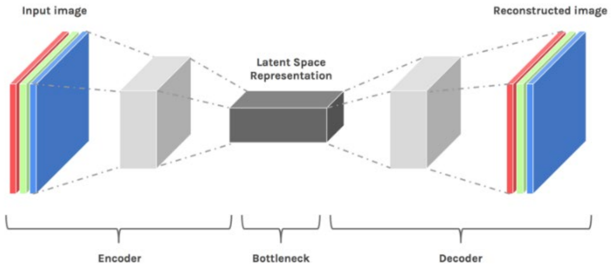
Convolutional NN (CNN) Example



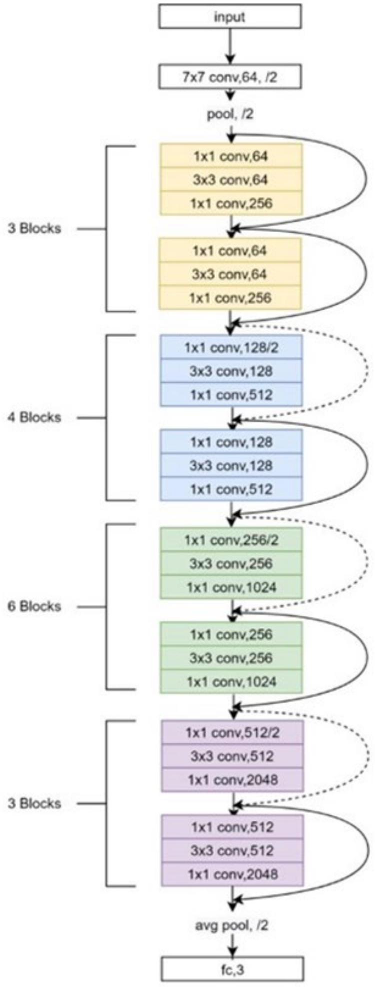
Transformer Example



U-Net Example



Autoencoder Example

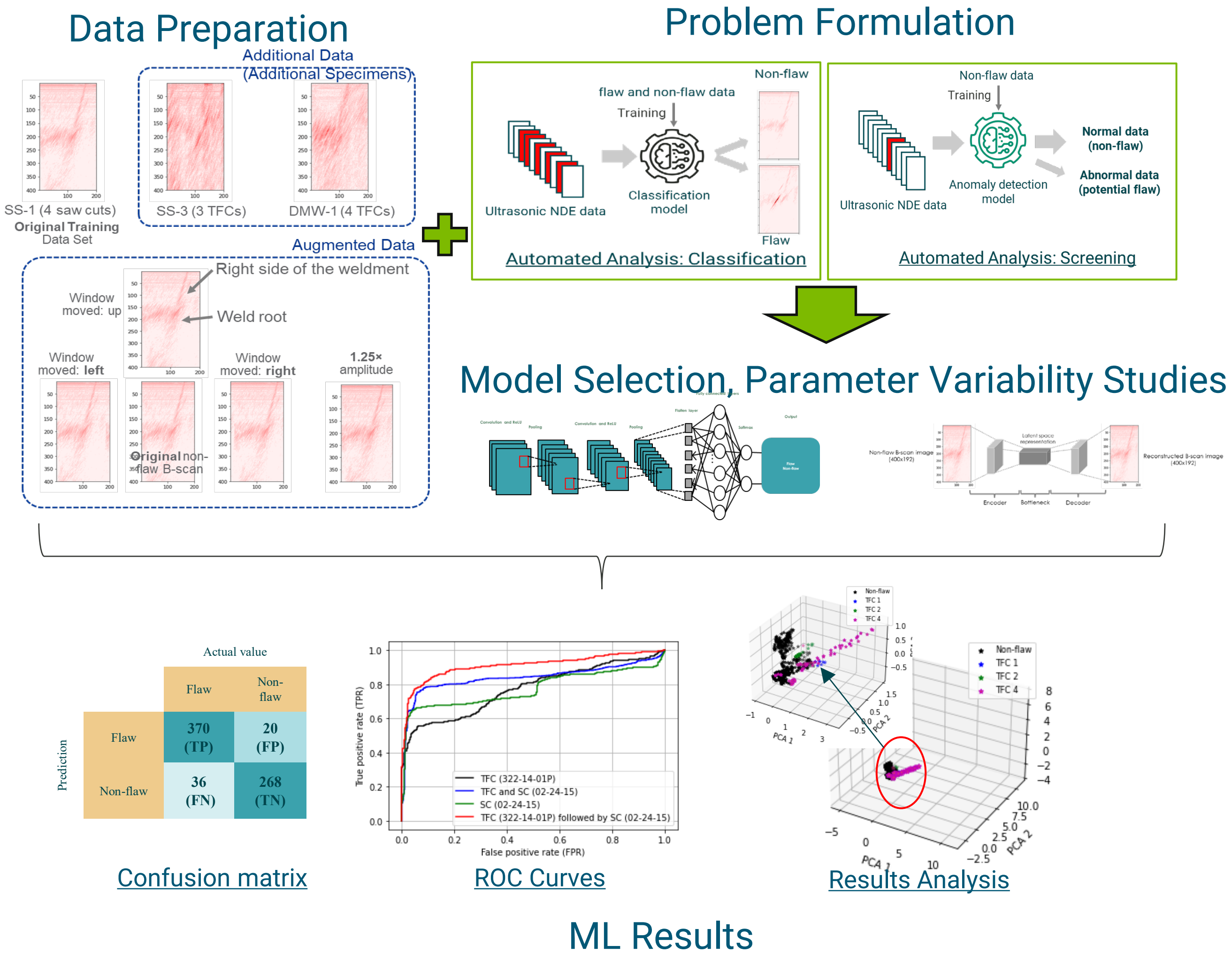


Resnet50

Evaluation approach uses a (growing) reference data set

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-036)	DMW pipe	1	TFC	63.0	58.0%
		2	TFC	72.4	95.0%
		3	TFC	40.1	35.0%
G(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%
H(9C-023)	DMW pipe	1	TFC	70.0	33.8%
		2	TFC	51.1	18.6%
		3	TFC	70.0	23.9%
		4	TFC	57.4	11.3%
I (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%
J (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%

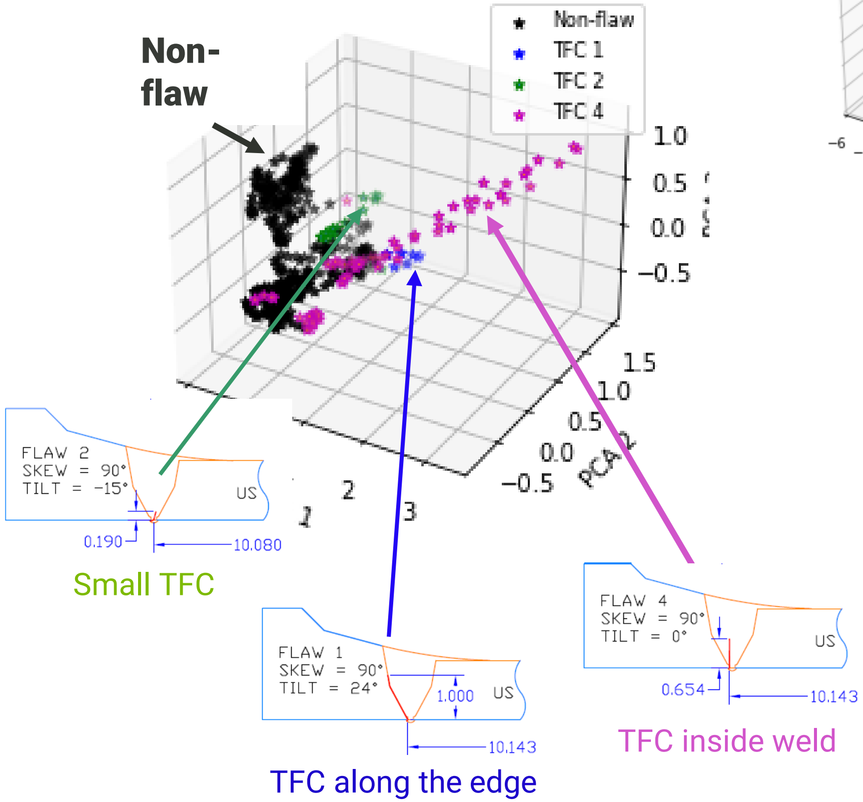
Flaws in Reference Dataset (41 total: 12 saw cuts, 23 thermal fatigue cracks, 6 EDM notches)



		Actual value	
		Flaw	Non-flaw
Prediction	Flaw	370 (TP)	20 (FP)
	Non-flaw	36 (FN)	268 (TN)

OAK RIDGE National Laboratory

Previous Analyses: ML performs best when training and test data are from similar distributions



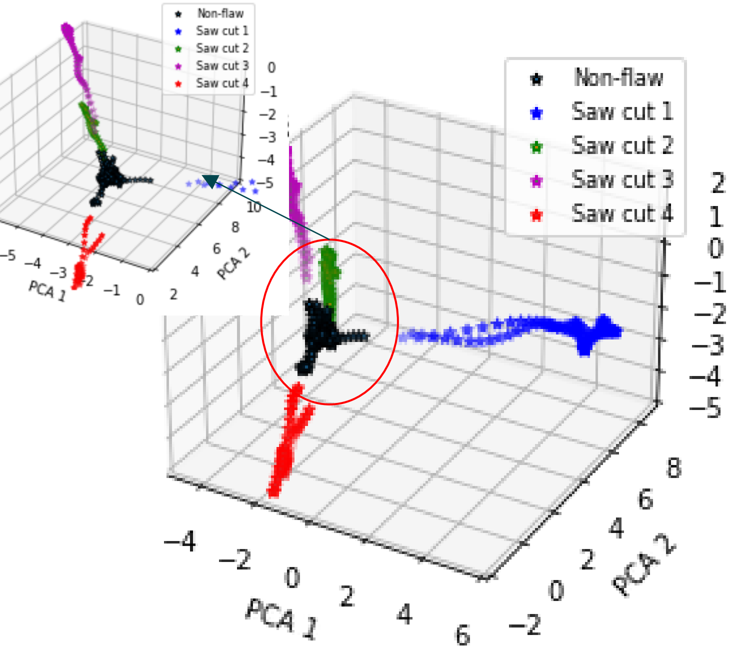
Non-flaw

Small TFC

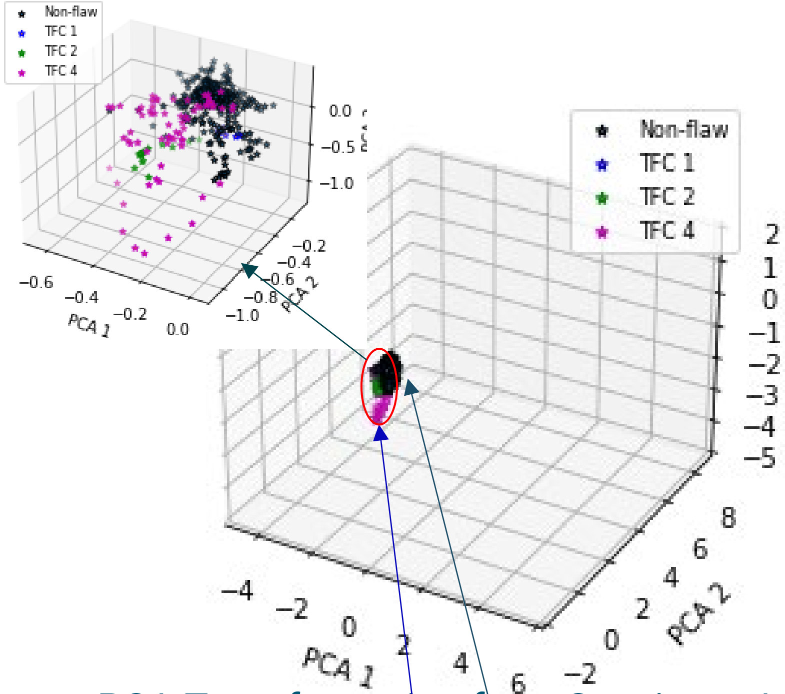
TFC along the edge

TFC inside weld

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˚)

Actual value

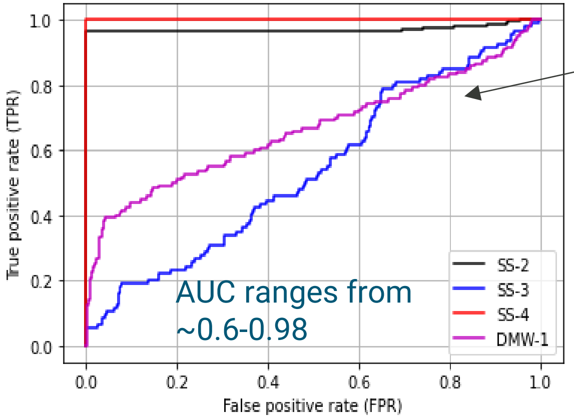
Prediction	Flaw	Non-flaw
Flaw	0 (TP)	0 (FP)
Non-flaw	128 (FN)	352 (TN)

Accuracy=0.73,
TPR=0, FPR=0

ML results for Specimen C (CNN model trained with Specimen A data)

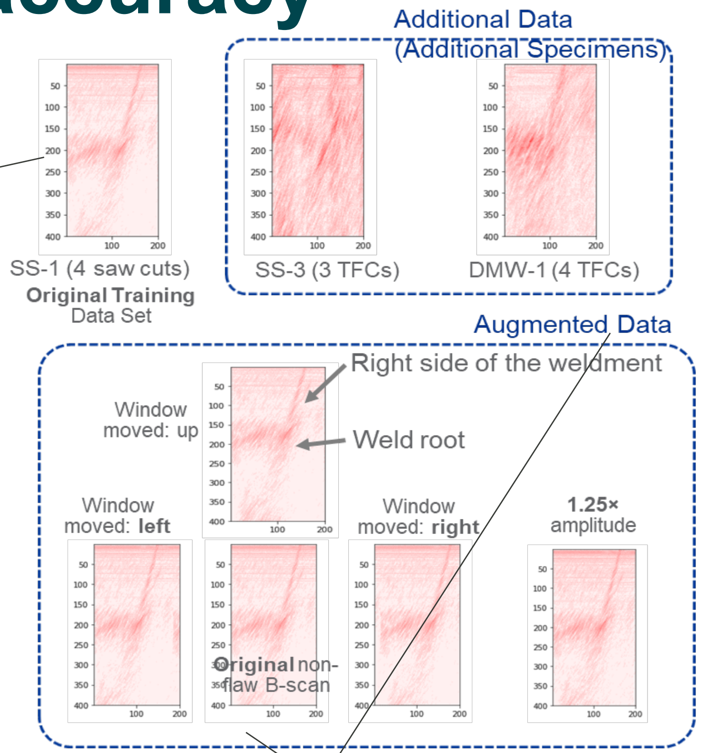
Principal Component Analysis (PCA) based on B-scan data (using three components)

Data augmentation can improve richness of training data for increased ML accuracy



AUC ranges from ~0.6-0.98

Anomaly Detection ROC Curves: before data augmentation



SS-1 (4 saw cuts)

Original Training Data Set

Additional Data (Additional Specimens)

SS-3 (3 TFCs)

DMW-1 (4 TFCs)

Augmented Data

Window moved: up

Right side of the weldment

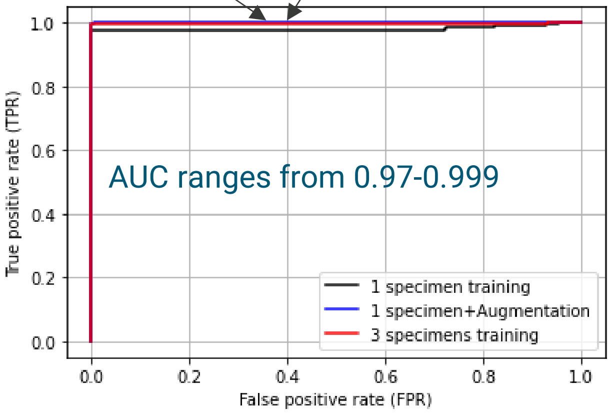
Weld root

Window moved: **left**

Original non-flaw B-scan

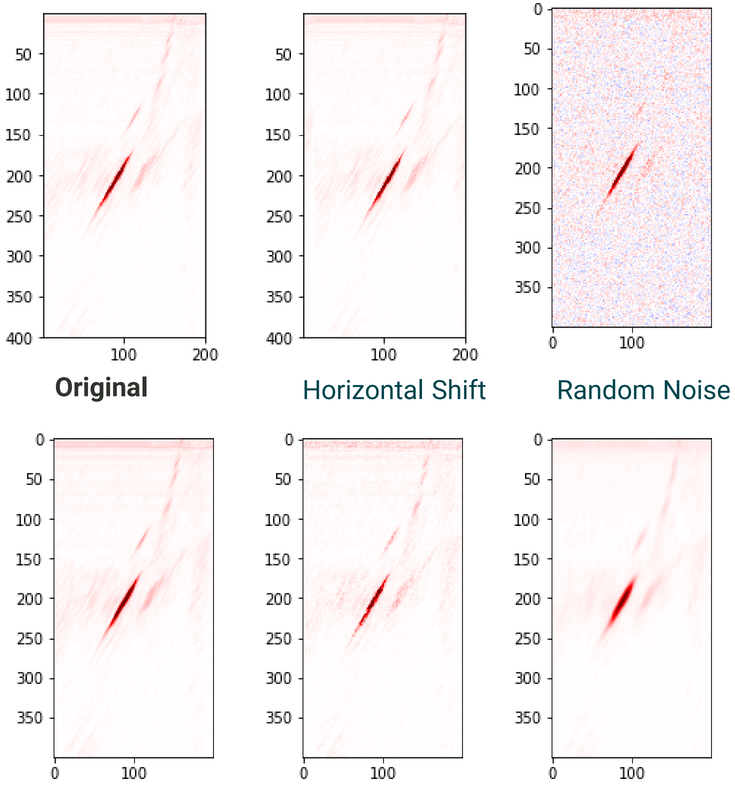
Window moved: **right**

1.25× amplitude



AUC ranges from 0.97-0.999

Anomaly Detection ROC Curves: after data augmentation



Original

Horizontal Shift

Random Noise

Gain Adjustment

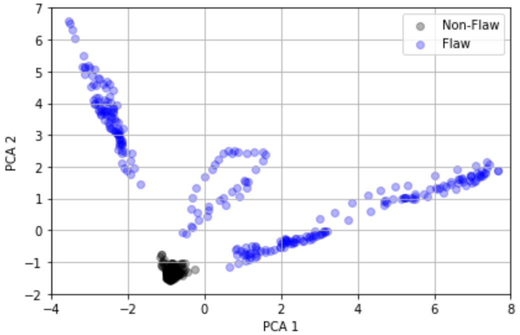
Speckle Noise

Filtering

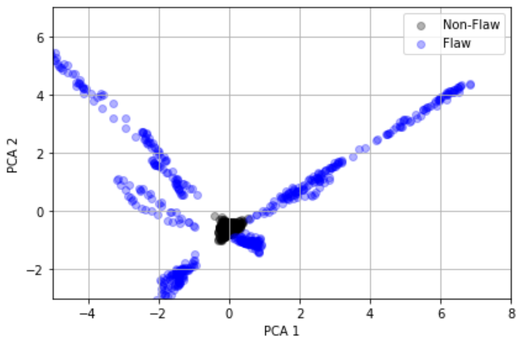
Other Impacts of Data Augmentation Being Evaluated

Data augmentation appears to increase data diversity

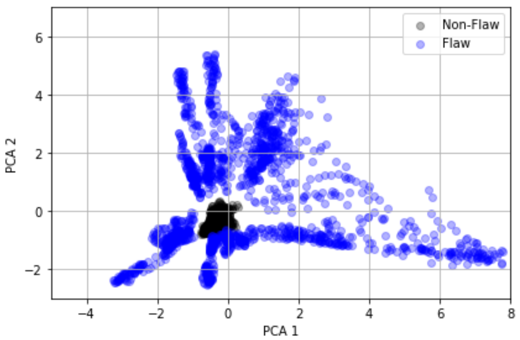
PCA, **Specimen D** (SwRI 45˚), original data



PCA, **Specimen D** original data + 1 augmentation data

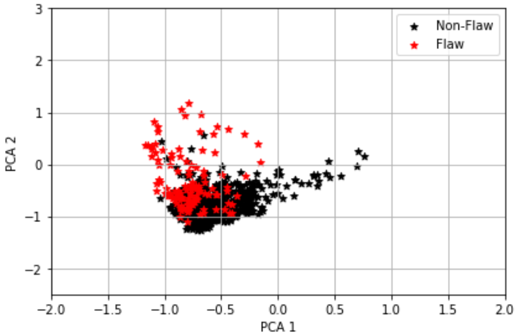


PCA, **Specimen D** original data + 5 augmentation data



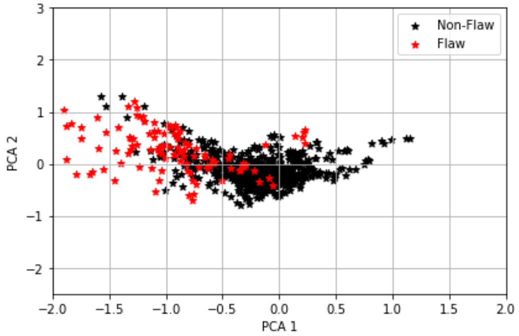
Flaw data diversity increases

PCA model fitted applied to **Specimen E**



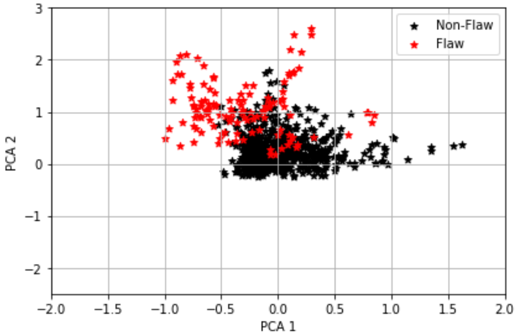
TPR=0.52, FPR=0.1

PCA model fitted applied to **Specimen E**



TPR=0.98, FPR=0.53

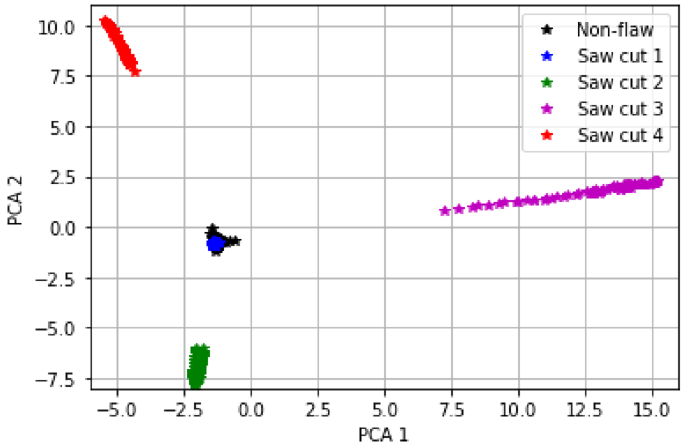
PCA model fitted applied to **Specimen E**



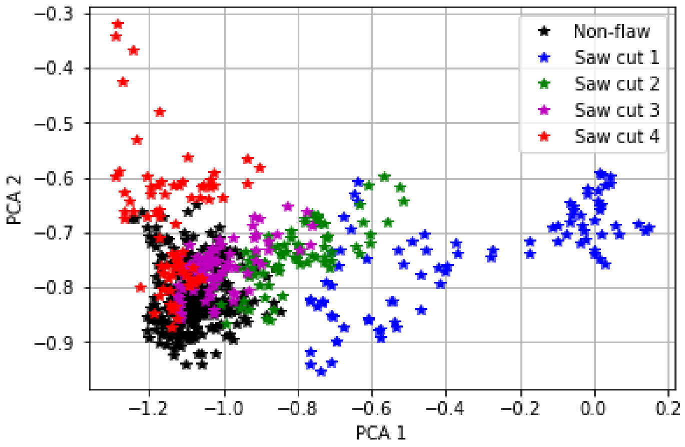
TPR=0.88, FPR=0.23

Test performance improves

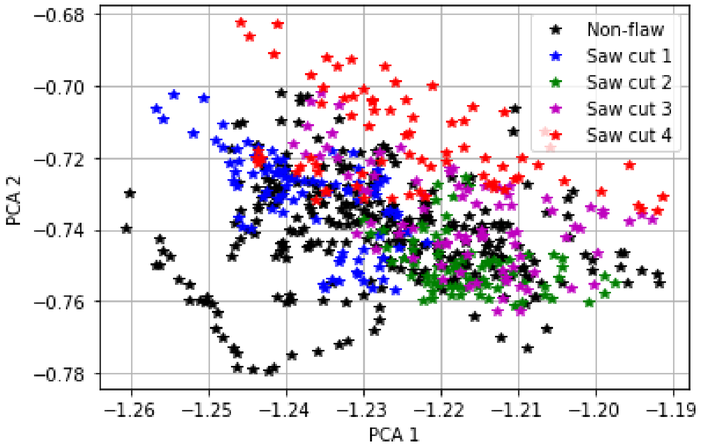
Data representativeness is equally important, even with data augmentation



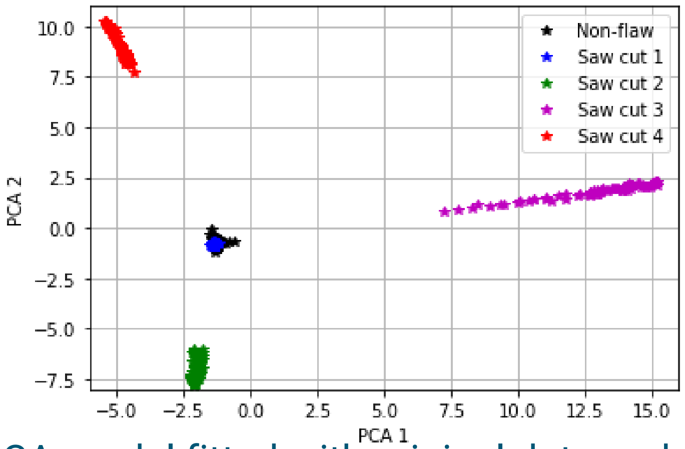
PCA, Specimen B (SwRI 45˚), original data



PCA model fitted with original data and applied to **augmented data (vertical shift)**



PCA model fitted with original data and applied to **augmented data (Horizontal flip)**

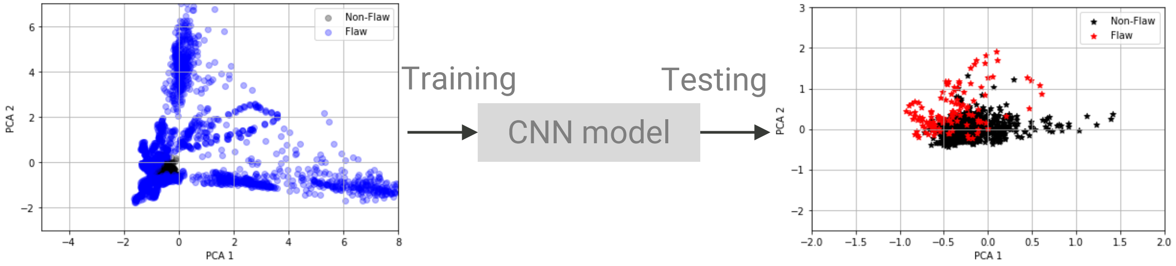


PCA model fitted with original data and applied to **augmented data (2% Gaussian noise)**

Metrics: ROC Curves and TPR/FPR

PCA, Specimen D original data
+7 augmentation data

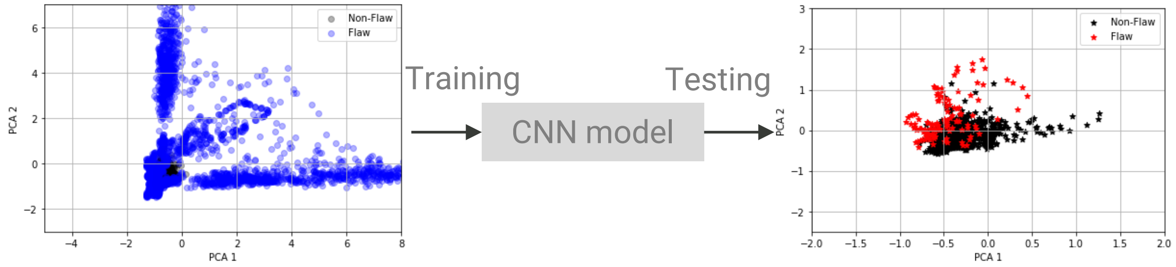
PCA model fitted applied to
Specimen E:



Training → CNN model → Testing

PCA, Specimen D original data
+9 augmentation data

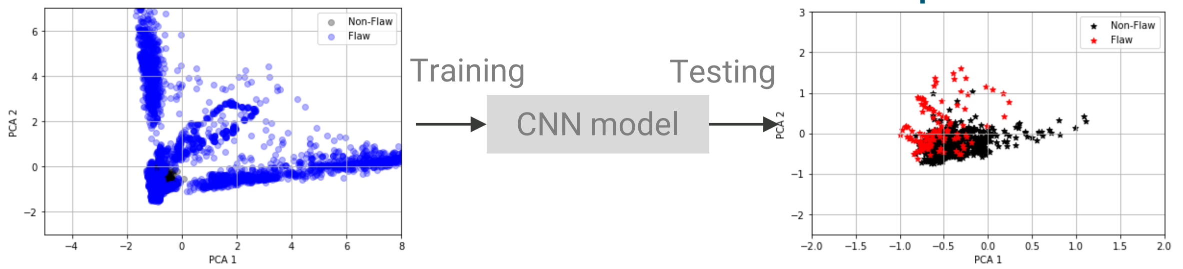
PCA model fitted applied to
Specimen E:



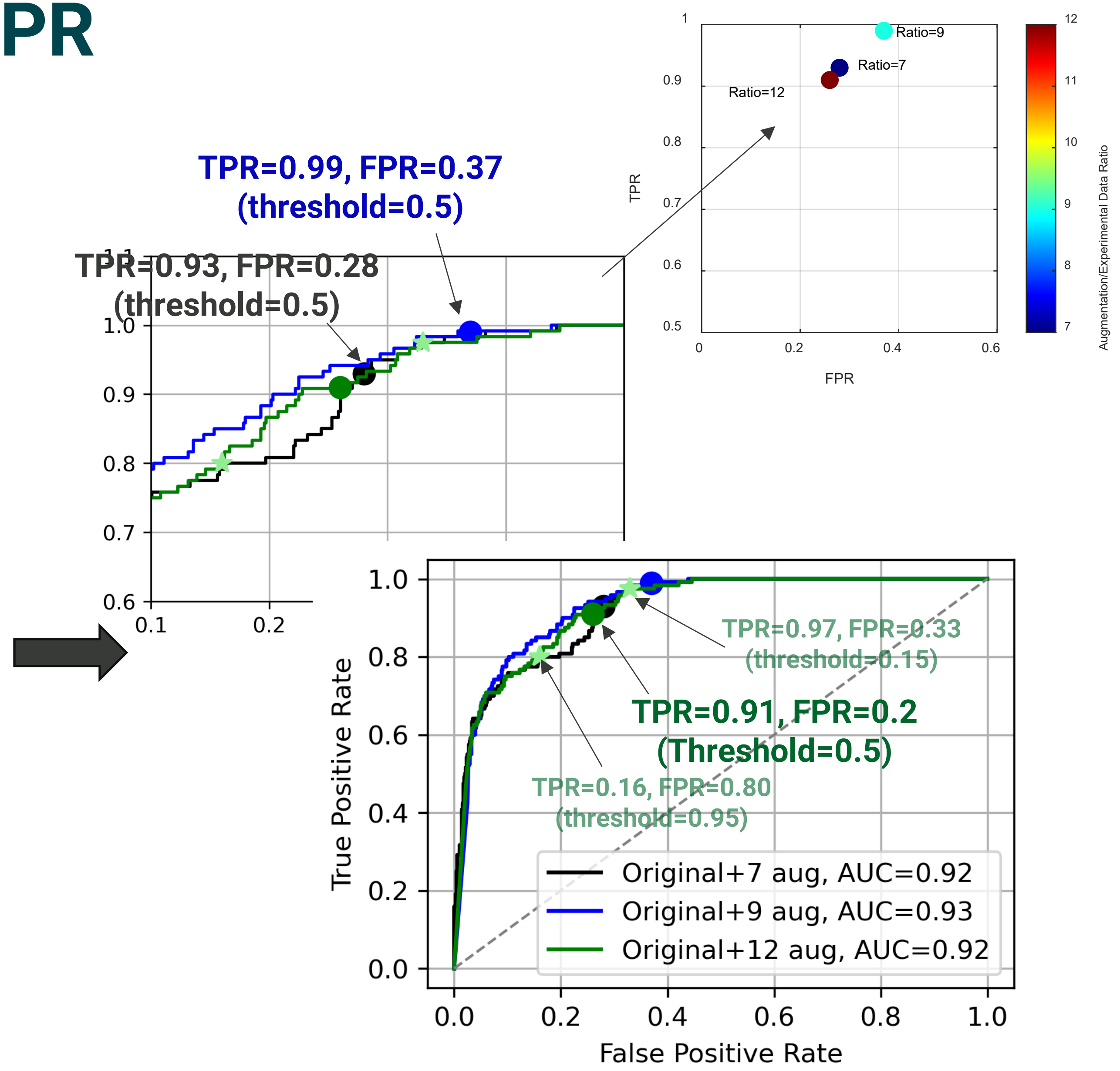
Training → CNN model → Testing

PCA, Specimen D original data
+12 augmentation data

PCA model fitted applied to
Specimen E:



Training → CNN model → Testing

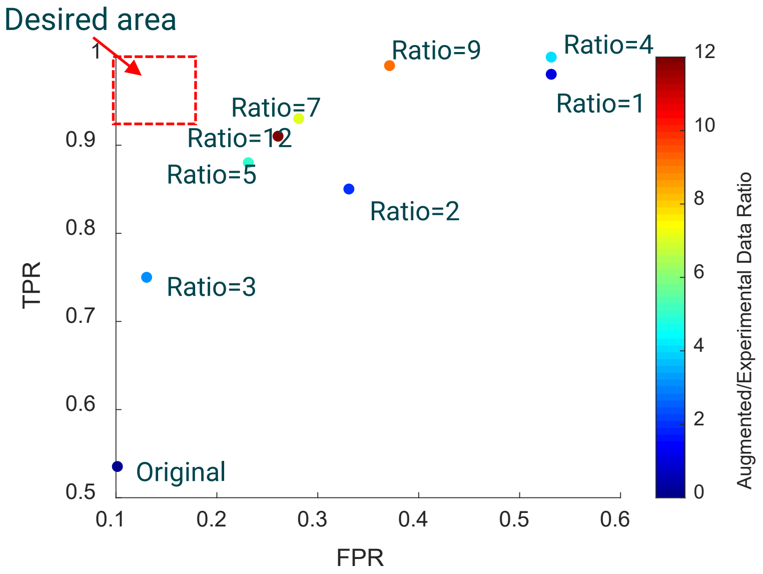


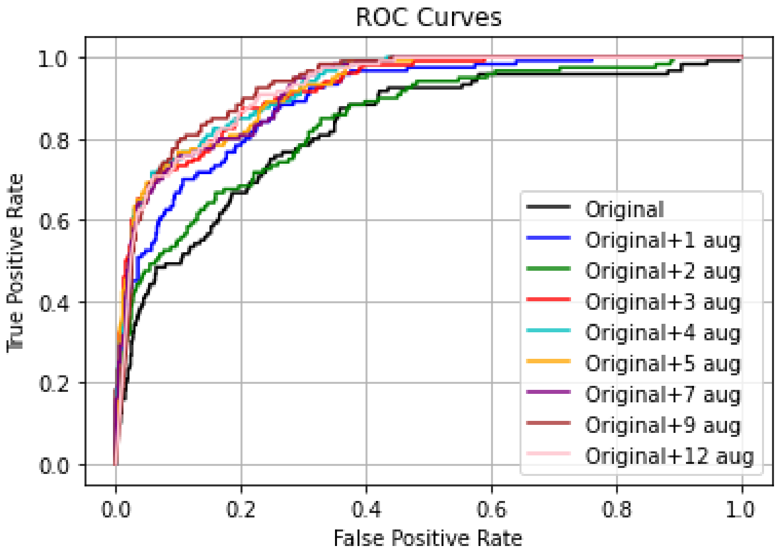
Receiver Operating Characteristic (ROC) curves (threshold: from 0 to 1)

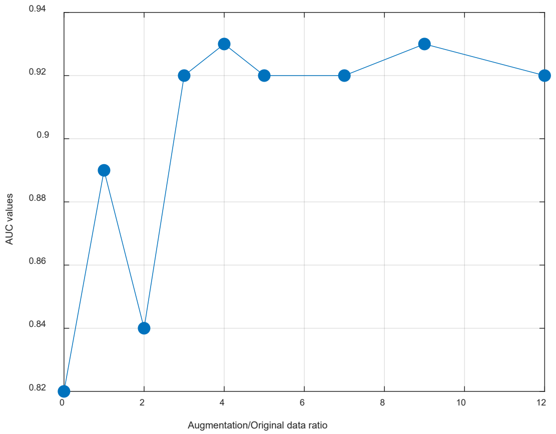
Increased Data Augmentation improves ML performance up to a point

Training with different **augmentation data size**

	Training (original + augmentation data)	**Testing**	**Original /Augmentation ratio**
0	02-24-15 (specimen D)	8C-032 (E)	N/A
1	02-24-15 + 1 aug	8C-032	682/682=1
2	02-24-15 + 2 augs	8C-032	682/1364=1:2
3	02-24-15 + 3 augs	8C-032	682/2046=1:3
4	02-24-15 + 4 augs	8C-032	682/2728=1:4
6	02-24-15 + 5 augs	8C-032	682/3410=1:5
7	02-24-15 + 7 augs	8C-032	682/4774=1:7
8	02-24-15 + 9 augs	8C-032	682/6138=1:9
9	02-24-15 + all 12 augs	8C-032	682/8184=1:12





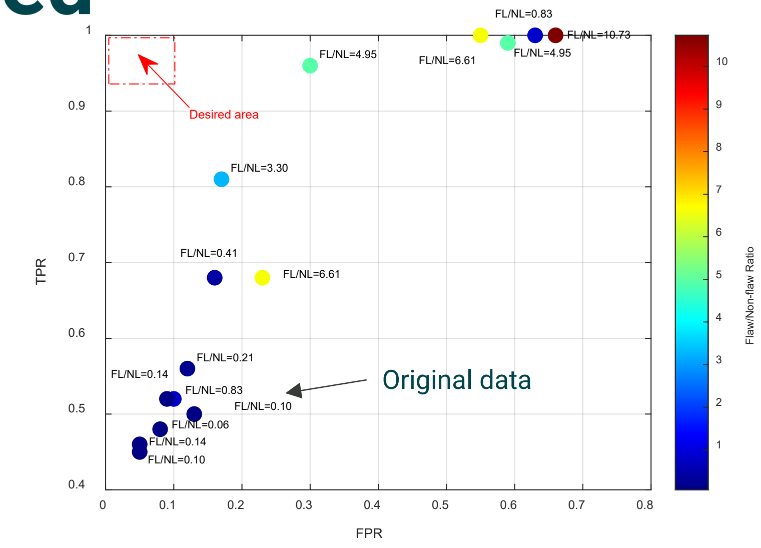


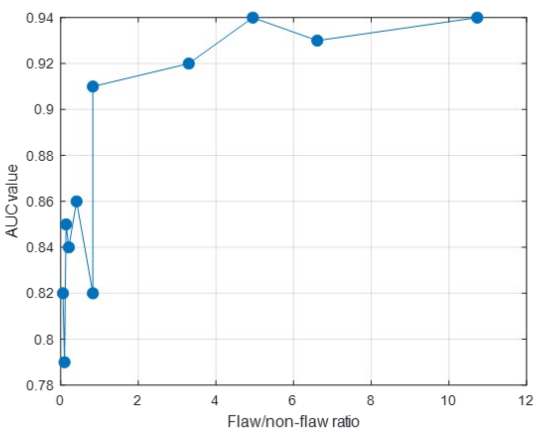
Area under curve (AUC) vs. Augmentation Level

Data imbalance during augmentation influences detection performance but it’s complicated

Training with different **flaw/non-flaw ratio**

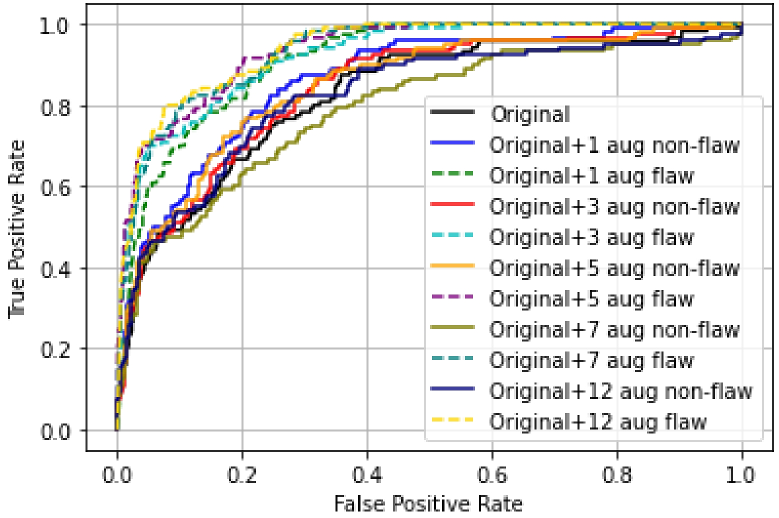
	Training (original + augmented flaw or non-flaw)	Testing	Flaw/Non-flaw Ratio
0	02-24-15 (specimen D)	8C-032 (E)	270/327=0.83
1	02-24-15+1 aug non-flaw only	8C-032	270/654=0.415
2	02-24-15+1 aug flaw only	8C-032	540/654=1.65
3	02-24-15+3 augs non-flaw only	8C-032	270/1308=0.21
4	02-24-15+3 augs flaw only	8C-032	1080/327=3.3
5	02-24-15+5 augs non-flaw only	8C-032	270/1962=0.14
6	02-24-15+5 augs flaw only	8C-032	1620/327=4.95
7	02-24-15+7 augs non-flaw only	8C-032	270/2616=0.1
8	02-24-15+7 augs flaw only	8C-032	2160/327=6.61
9	02-24-15+all 12 augs, non-flaw only	8C-032	270/4251=0.06
10	02-24-15+all 12 augs, flaw only	8C-032	3510/327=10.73





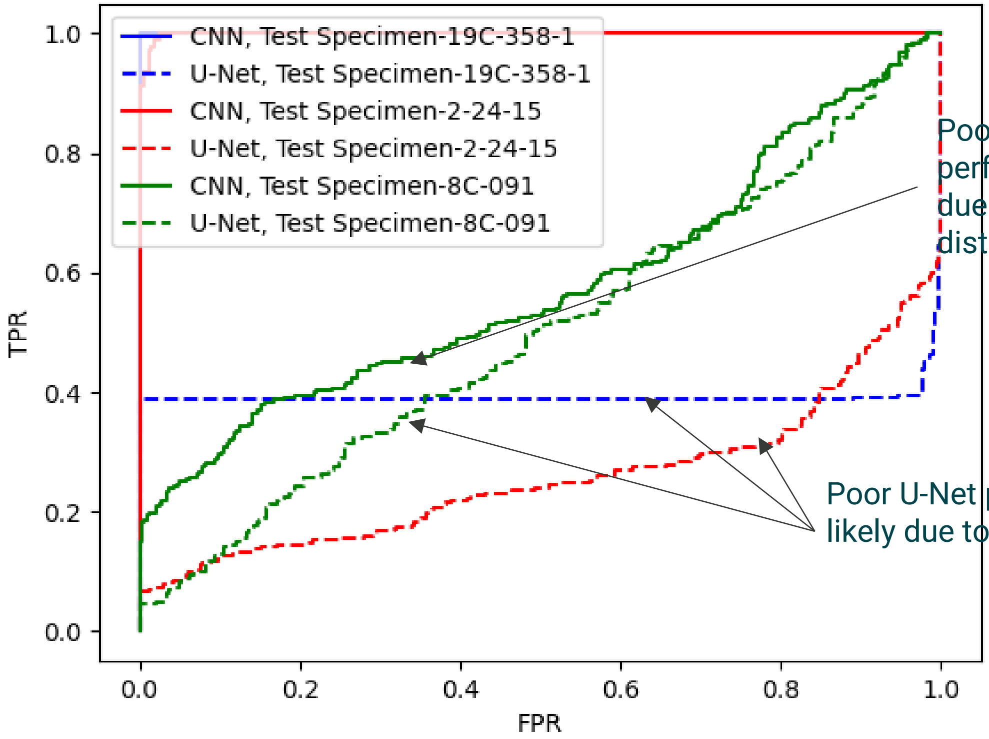
Area under curve (AUC) vs. flaw/non-flaw ratio

ROC Curves



OAK RIDGE National Laboratory

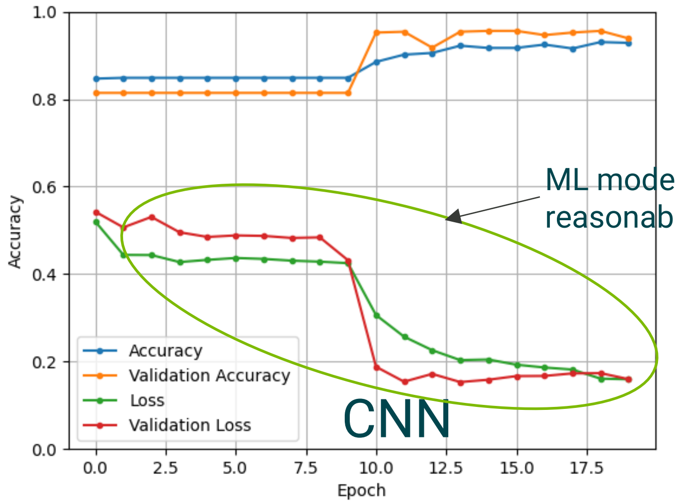
Comparison of ML Models highlights need to match model complexity with available data



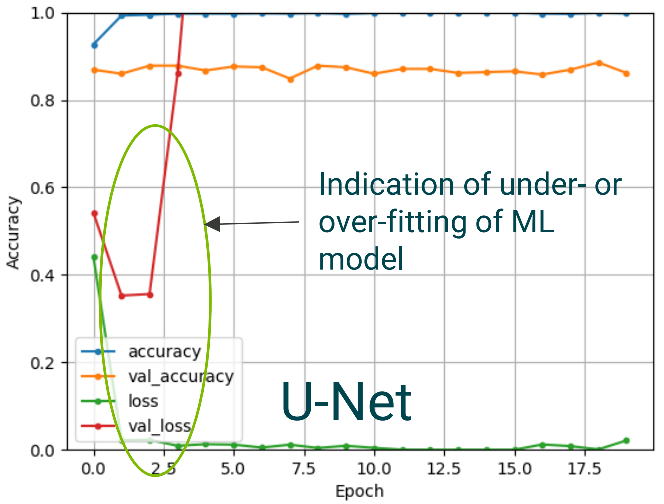
Poor CNN performance likely due to out of distribution test data

Poor U-Net performance likely due to under-fitting

Comparison of ROC curves for the CNN and U-Net classification models, which were trained on specimens 322-14-01P, 8C-032, and 8C-032 with data augmentation (signal amplitude only)



ML model – apparent reasonable fit to data

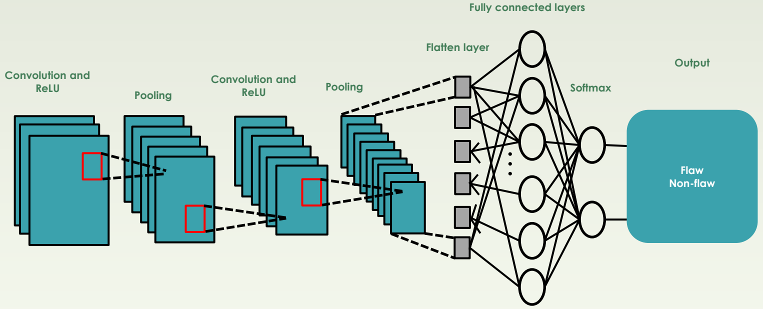


Indication of under- or over-fitting of ML model

OAK RIDGE
National Laboratory

Examples of performance for different ML models

Supervised learning
CNN classification model

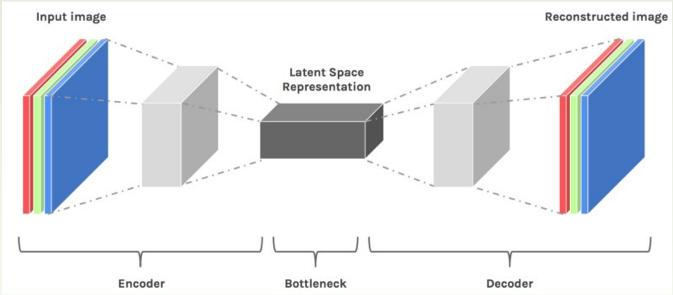


Training: specimen A
Testing: specimen D

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

TFR=0.97, FPR=0

Supervised learning
Autoencoder Anomaly Detection model

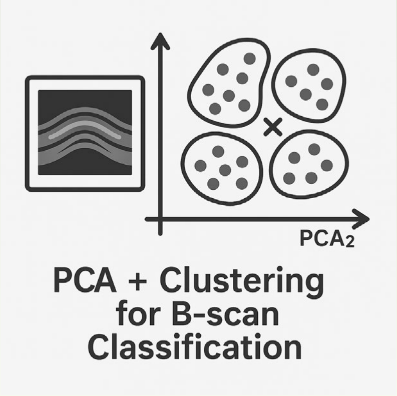


Training: specimen A
Testing: specimen D

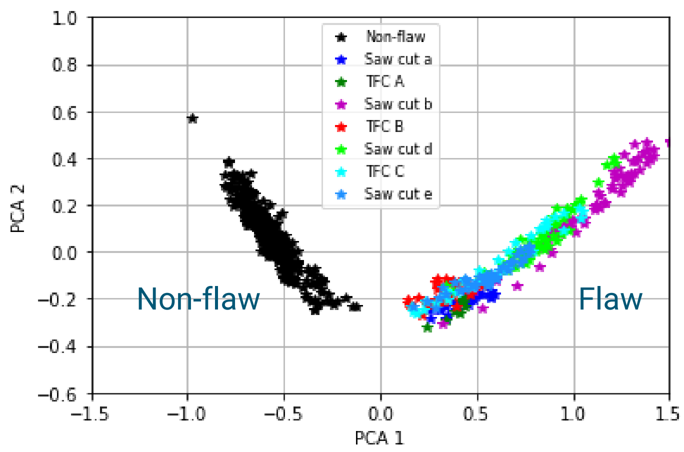
		Actual value	
		Flaw	Non-flaw
Prediction	Flaw	**268 (TP)**	**183 (FP)**
	Non-flaw	**1 (FN)**	**144 (TN)**

TFR=0.99, FPR=0.56

Unsupervised learning
(feature extraction+ PCA + clustering)



Specimen D

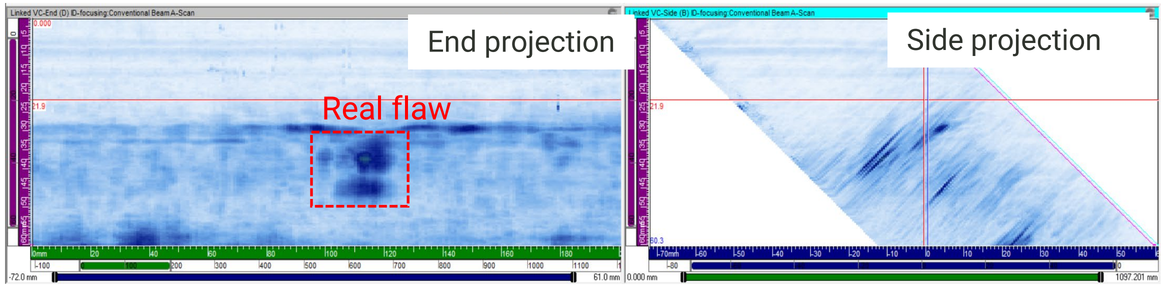


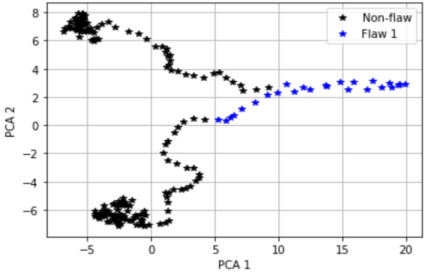
TFR=1, FPR=0

Ongoing: Virtual Flaws for Data Augmentation (specimens 11&13)

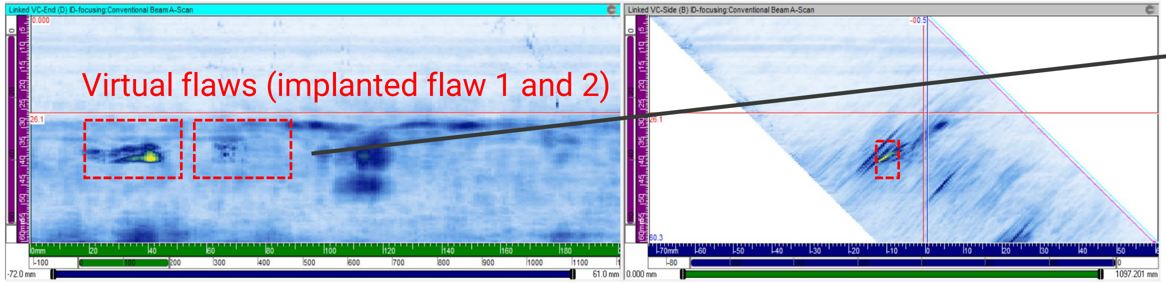
PCA plots based on B-scans

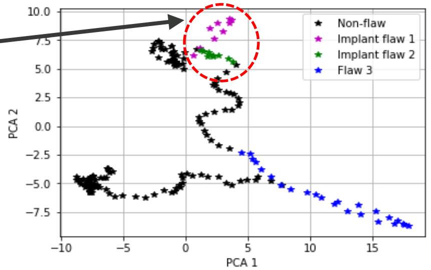
DMW (8C-032), TFC
(L: 26 mm, H: 8 mm)



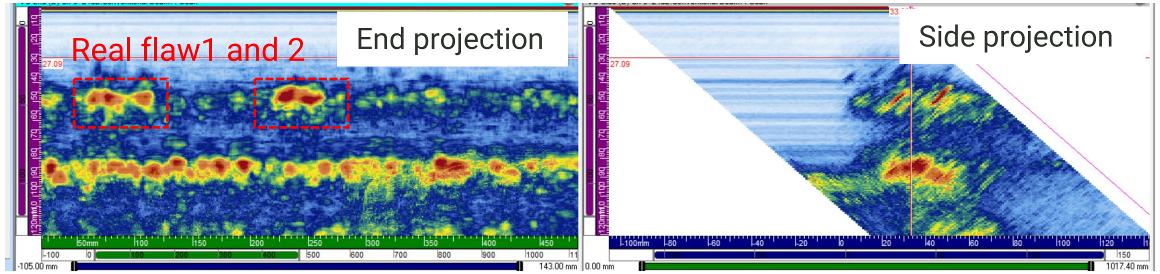


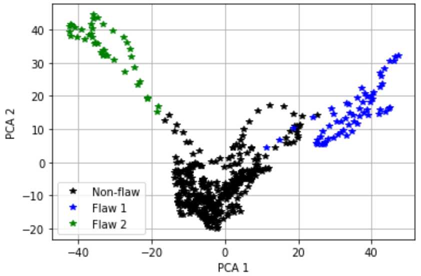
Virtual flaw 1
(L: 29 mm, H: 7 mm)
Virtual flaw 2
(L:11 mm, H: 7 mm)



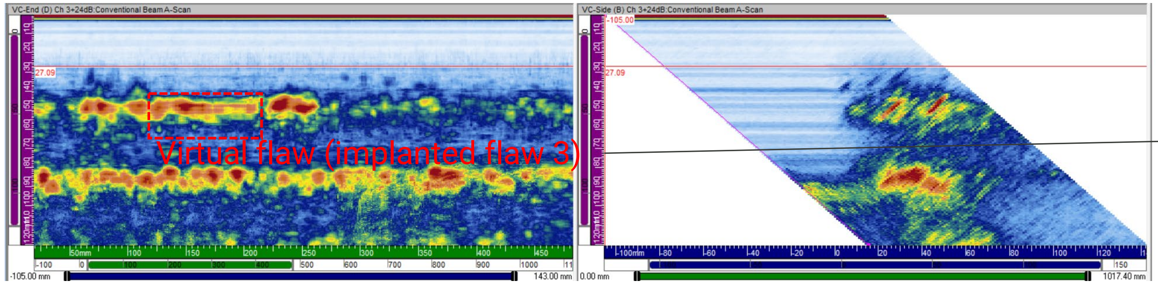


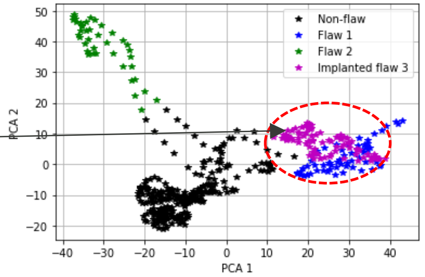
(707P1)
(L: 62 mm, H: 19 mm
L: 43 mm, H: 12mm)



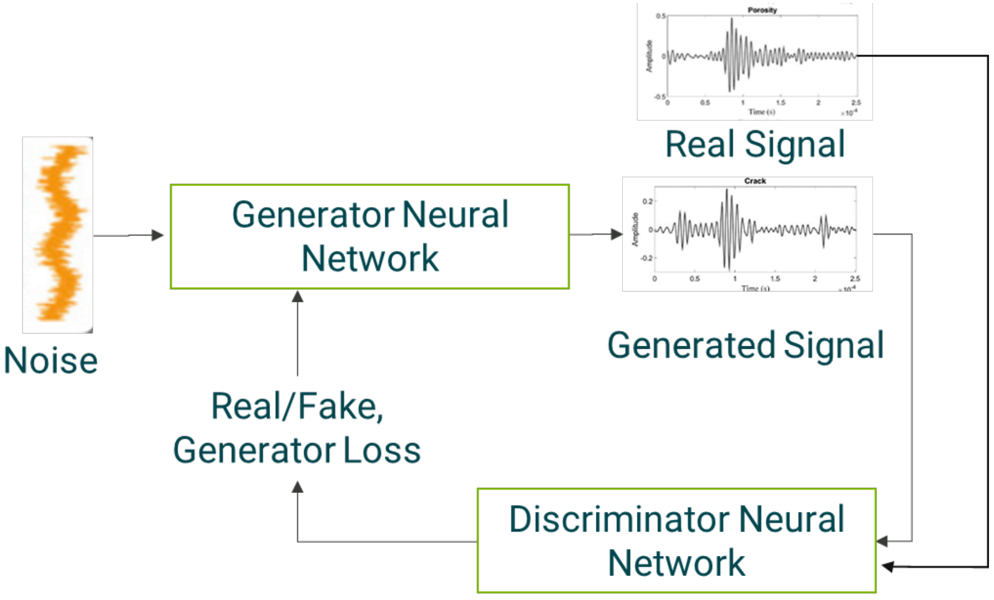


Virtual flaw 1
(L: 91 mm, H: 12 mm)

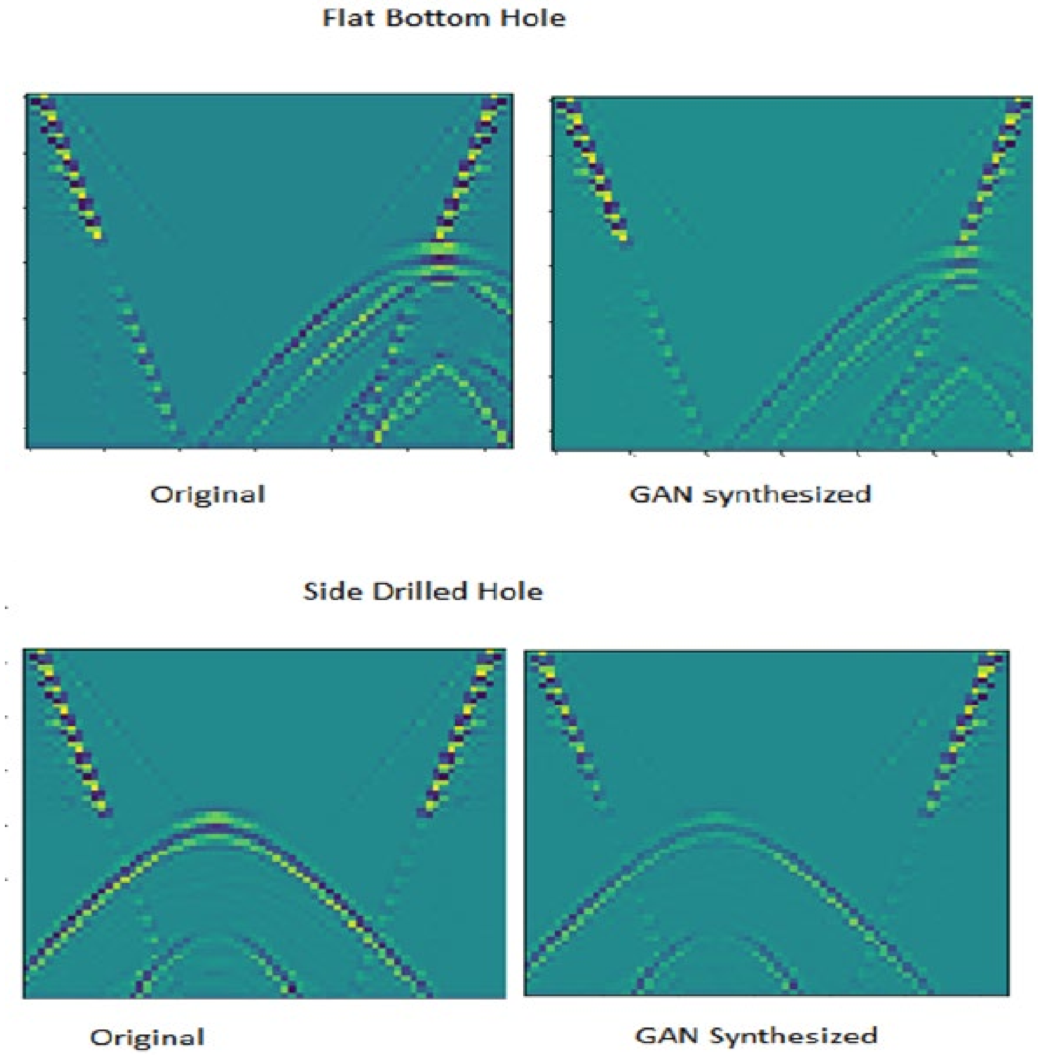




Deep Learning for Data Generation



Generative Adversarial Network (GAN)



Example of synthetic vs. real ultrasonic B-Scans generated by a GAN**.

**K. Virupakshappa and E. Oruklu, in *Proc. 2020 IEEE Int'l. Ultra. Symp.*, Las Vegas. doi: 10.1109/IUS46767.2020.9251382.

Ongoing Verification & Validation Activities

Model Correctness & Sensitivity

- Establish a fixed baseline ML model and training protocol
- Verify implementation correctness via: Multiple random seeds, data splits, and retraining runs
- Sensitivity studies to model type and hyperparameters
- Identify dominant factors influencing ML reliability

Benchmarking Against Qualified Human Analysts

- Statistically compare ML performance to expert human analysis for:
  - Detection, discrimination, and characterization
- Evaluate ML consistency relative to inter-analyst variability
- Establish ML performance context relative to Code-based human benchmarks

Performance Metrics for ML V&V

- Evaluate and verify candidate metrics for:
  - Detection (e.g., POD, AUC, recall at fixed false-call rate)
  - Classification (confusion matrix, class-wise accuracy, F1)
  - Characterization (error, bias, uncertainty bounds)
- Select metrics that are robust, interpretable, and aligned with inspection decisions

Reproducibility & Requalification

- Quantify performance variability across training runs and datasets
- Define reproducibility metrics and acceptable bounds
- Establish triggers and scope for requalification as data, models, or procedures change

Summary

- ML models were evaluated with ultrasonic NDE data collected on weld mockups (austenitic and dissimilar metal welds).
- Results to date suggest that ML has the potential for supporting automated NDE data analysis if applied appropriately.
- Results indicate several factors, including data richness and representativeness, play a role in ML accuracy.
  - Mode converted responses and tip signals may contribute to multiple flaw calls
  - Material/weld noise may contribute to higher false call rate
  - Models and formulations need to be matched to available data sets
- Increasing the richness of training data using data augmentation techniques seems to improve ML performance, though attention needs to be paid to other factors such as representativeness and balance in data
- Assessment of other ML advances is ongoing

Future Work

- ML
  - Evaluating generative AI for training and test data generation.
  - Examining model explainability approaches.
- Evaluation
  - Other statistical metrics for ML performance, including POD.
- Qualification
  - Verification and validation criteria for ML qualification for NDE applications

Questions?