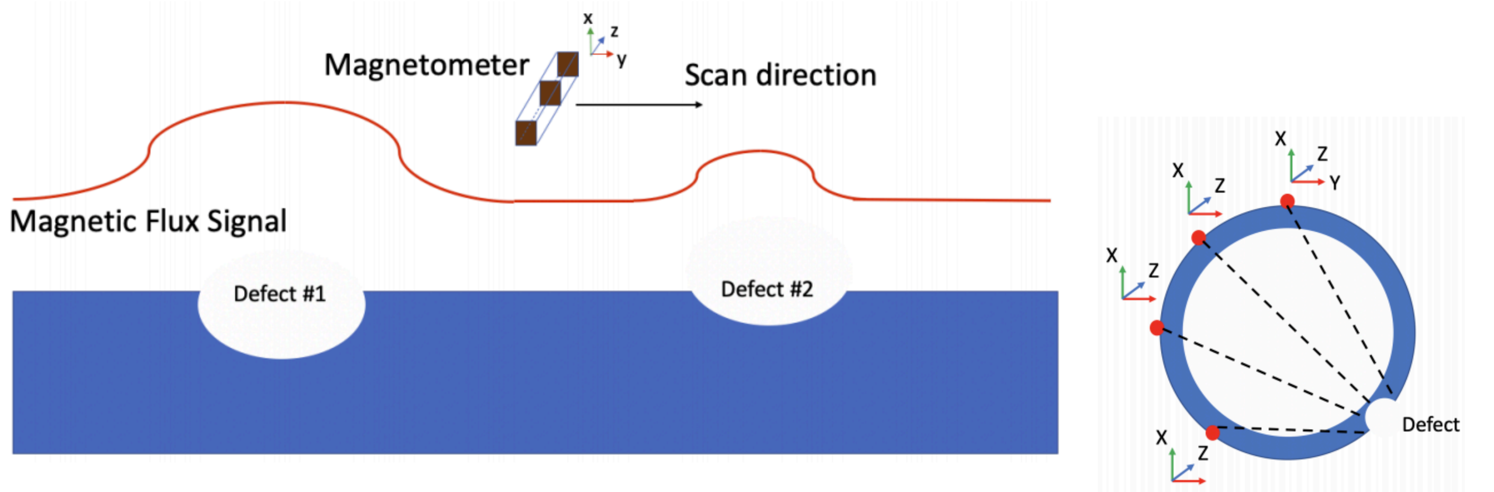
Machine Learning-based Anomaly Detection with Magnetic Data

ML4Eng Paper

Peetak P. Mitra, Denis Akhiyarov, Mauricio Araya-Polo, Daniel Byrd

Motivation

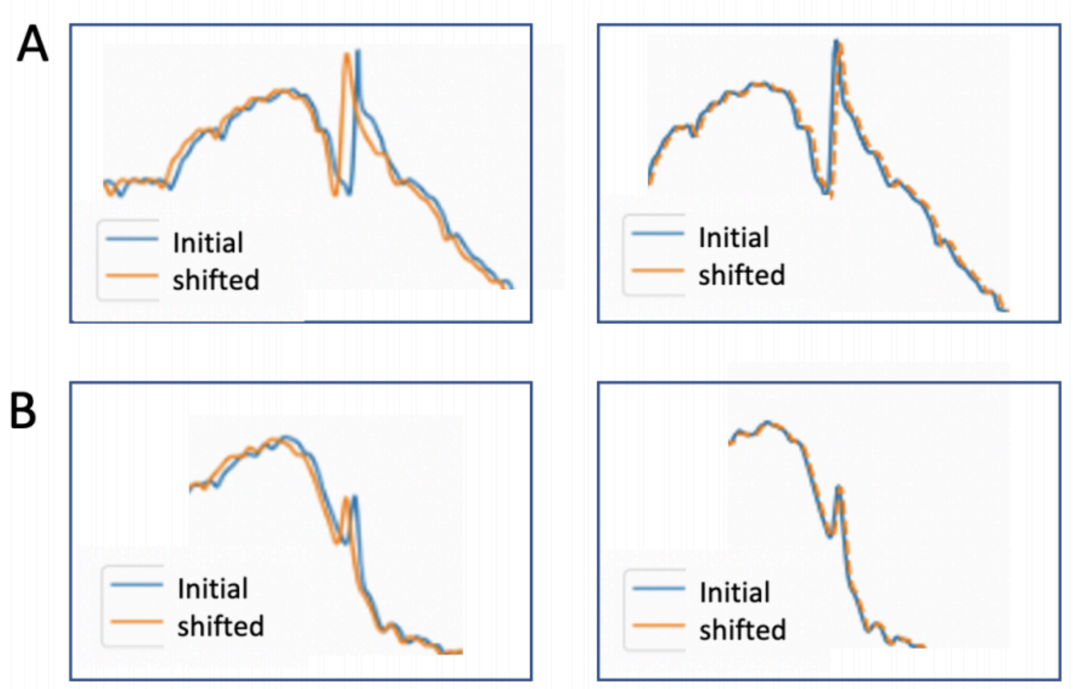
- Pipeline Integrity important to the energy industry
- Undetected defects can cause significant damage
- Intrusive methods cause operational challenges
- Non-intrusive magnetic methods like LSM are promising in detecting/characterizing pipeline defects
- Anomaly detection from multi-sensor, multi-alignment LSM data not trivial
- Study to explore Scalable ML methods for this task



Schematic of LSM technology showing data collection across multiple sensors, and gathers data in all three spatial directions making the data collection multi-modal in nature

Data and Preprocessing

- Multi-sensor LSM data are multi-modal, non-aligned sequences, that affects ML model predictions
- Fast Dynamic Time Warping algorithm re-aligns dataset with O(N) time and space complexity

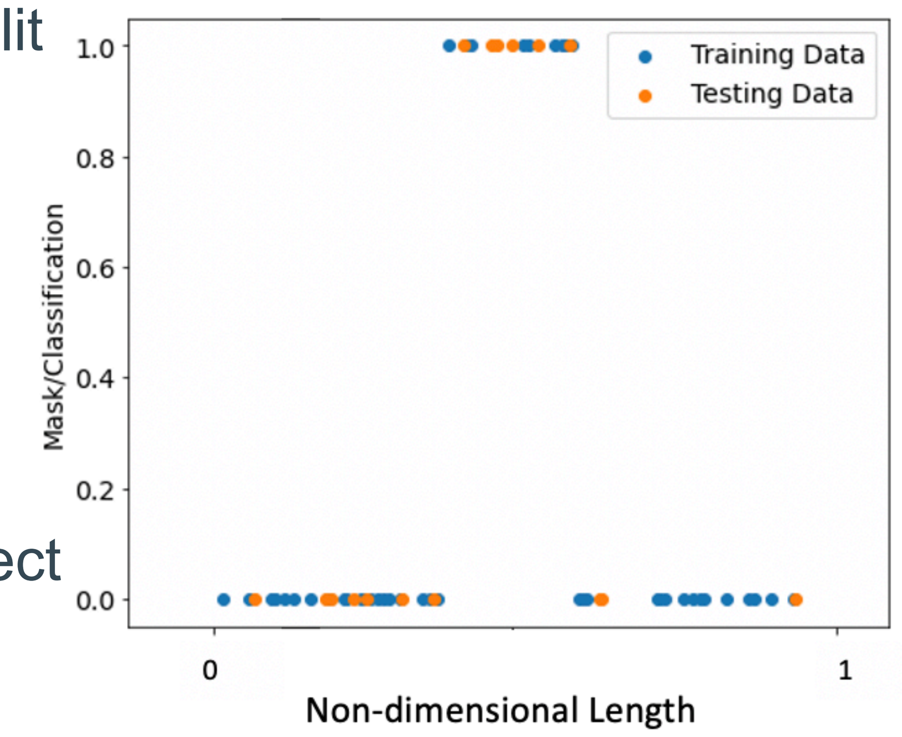


Before and after alignment snapshots of multi-sensor data

Defect	Location	Volume	Depth	Width
D1	2 ft	0.2	0.77	0.45
D2	76 ft	0.6	0.62	1

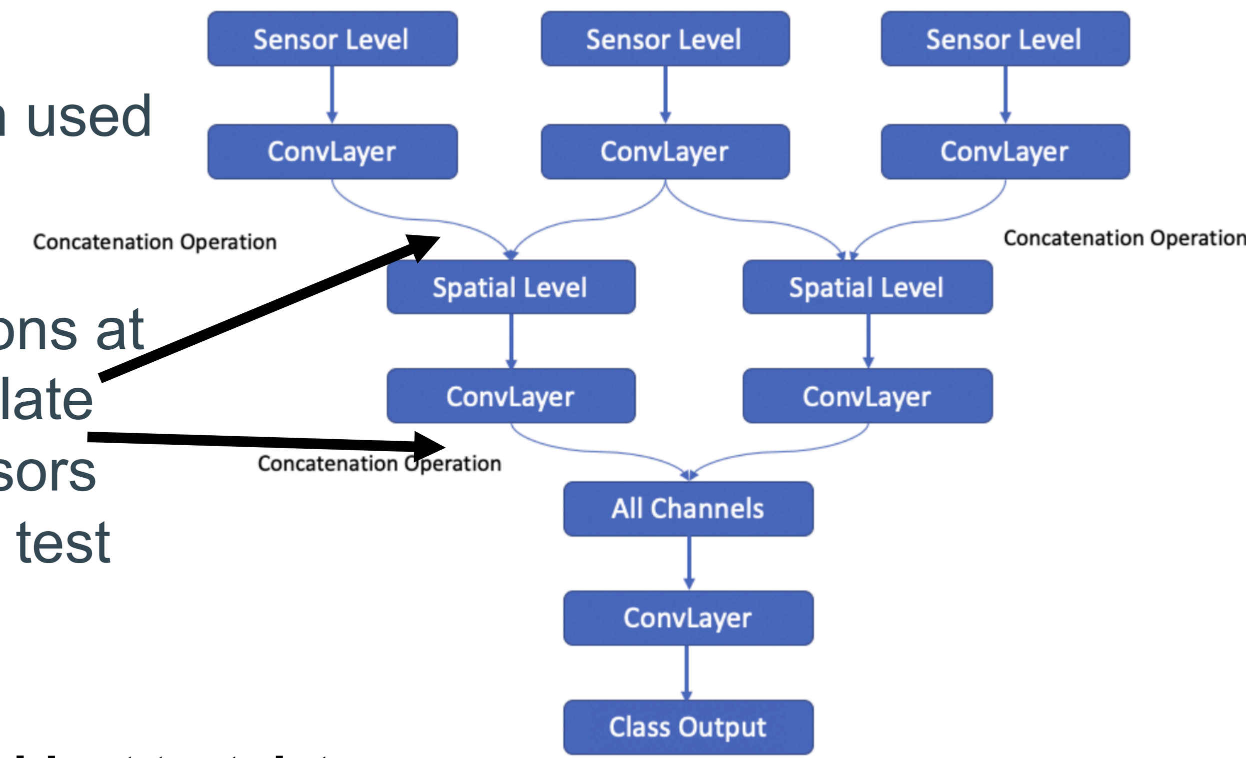
Customized 1D CNN for multi-output prediction

- Data leakage in random test/train split
- Masked defect regions with +/- 10 ft
- **Train samples:** 35000
- **Test samples**: 10000



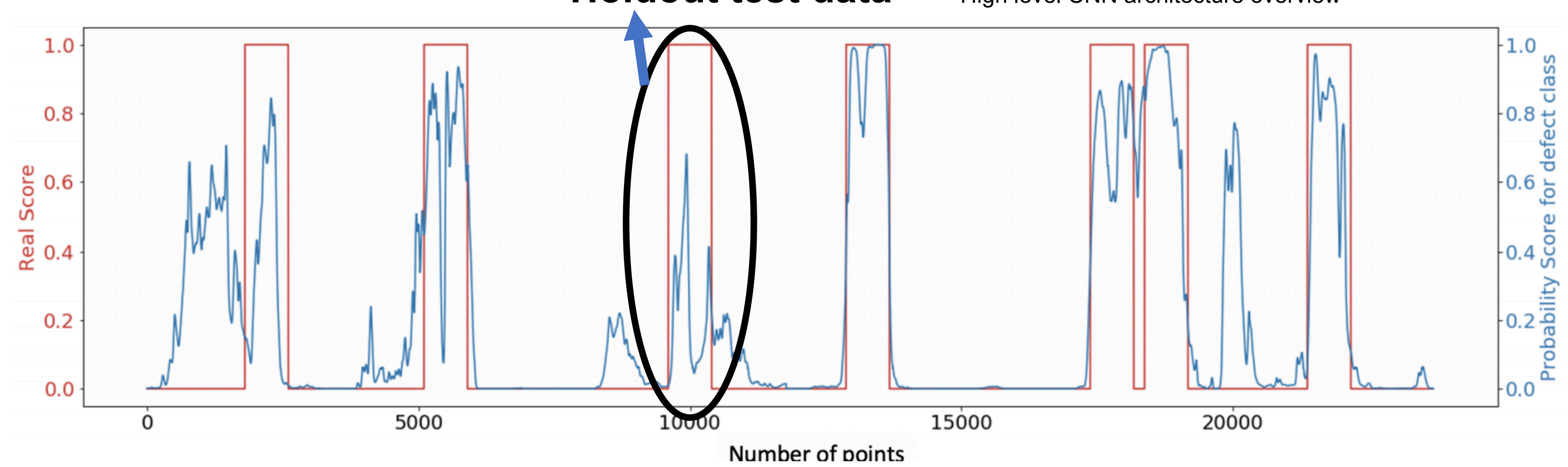
Close overlap of train/test data shows data leakage, which affects model predictions

- The "point-based" methods can detect defect, but not characterize them
- Sequence learning using CNN with 1D filters used to extract spatial features
- Multi-task classification used to characterize defect properties
- Concatenation operations at the spatial levels to isolate effects of different sensors
- One defect held out as test dataset



High level CNN architecture overview

ConvNet Performance



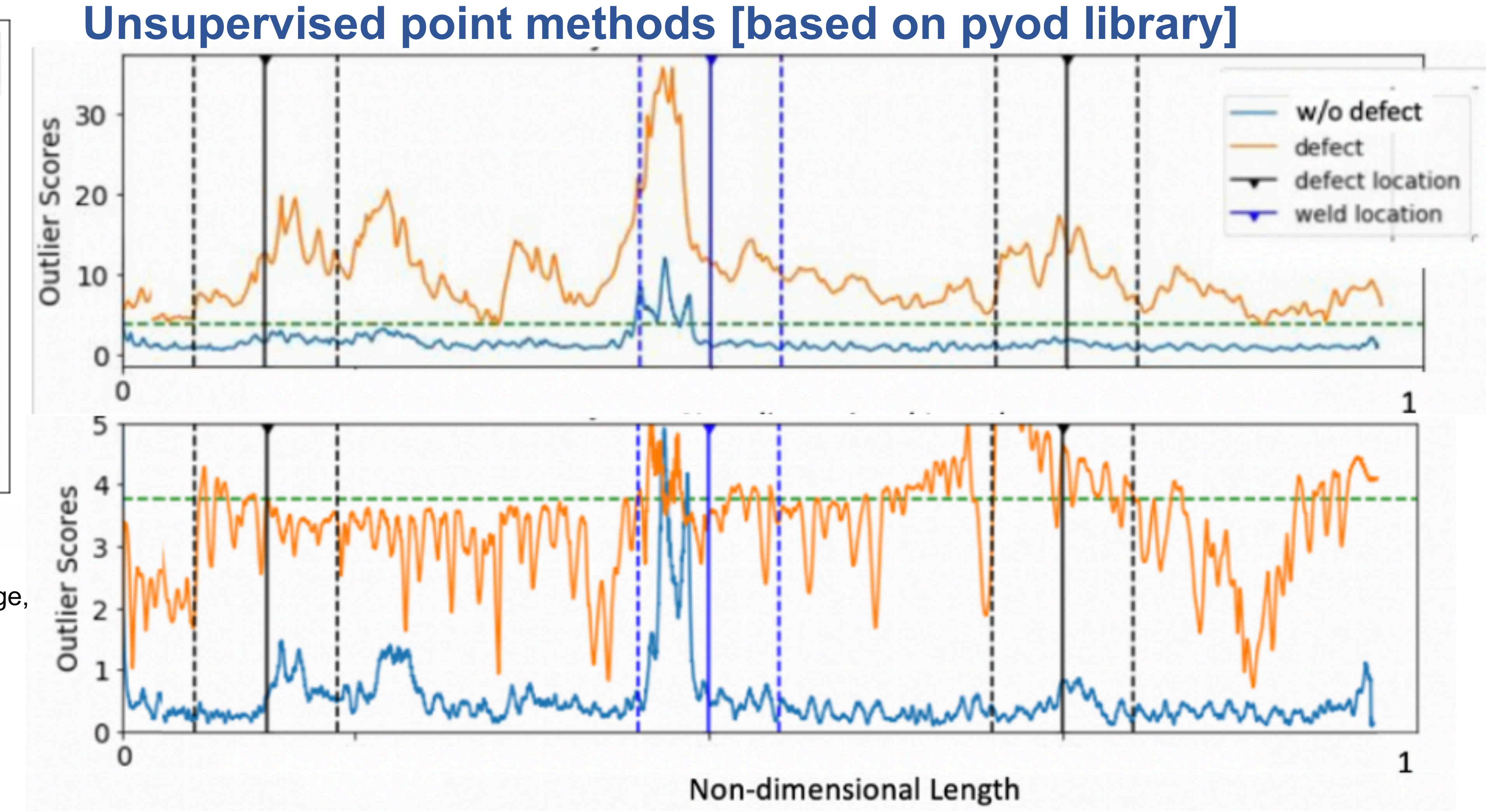
ConvNet performance on train/test data shows high probability scores within masked defect regions, including unseen test data

TL;DR

- Robust multi-sensor data alignment using FastDTW achieved
- Point-based supervised/unsupervised learning methods identify defects successfully.
- Slower methods sped up using RAPIDS-AI cuML library
- Multi-output CNN techniques are useful tools for characterizing defects
- Feasibility for field data explored and suitable methods identified

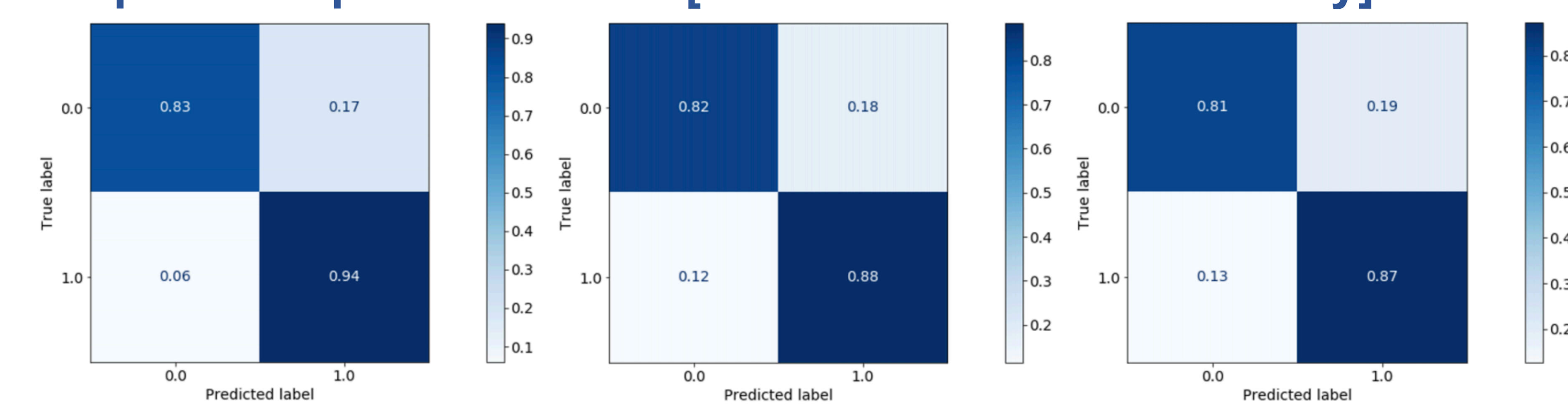
Off-shelf ML packages

Unsupervised point methods [based on pyod library]



k-NN outlier scores (in orange) higher in the defect regions, marked by dashed vertical black lines with few F.P.

Supervised point methods [based on scikit-learn library]



SVC · Decision Trees · MLP Classifier

Training time

- N = 10000 points
- All times in seconds
- SVC is slowest!

Algorithm	N	10*N	100*N
k-NN	1.47	11.2	140
SVC [rbf]	8.76	751	18274
Decision Trees	0.33	3.86	131
MLP Classifier	7.9	69	772

Speed up of SVC [RBF kernel] using RAPIDS-AI cuML library

Data points	scikit–learn	RAPIDS-AI	Speed up
10000	8.76	2.90	3
100000	751	3.75	200
1000000	18274	98	186

References

- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit- learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
- Yue Zhao, Zain Nasrullah, and Zheng Li. Pyod: A python toolbox for scalable outlier detection. *arXiv preprint arXiv:1901.01588*, 2019.

Acknowledgements

Deepwater R&D team for steering and domain expertise, as well as HPC team at TOTAL E&P