

Towards Improved Hydrocarbon Soil Assessment: The Application of Mid-Infrared Spectroscopy and Binary Classification Techniques

Deeksha Beniwal*, Radha Krishnan Nachimuthu, Chaya Smith,
Georgios Tsiminis, Sean Manning

2024 Chlorinated Conference

Session: Data Analytics: Use of Machine Learning and Artificial Intelligence Tools for Improved Analysis, Optimization and Decision Making

Motivation

- Rapid on-site assessment is needed for efficient remediation.
- Traditional lab tests are expensive, slow and resource-intensive.
- RemScan is a fast, cost-effective measurement solution.
- Extensive work done to calibrate the instrument.
- New calibration method being developed for Remscan.
- Improve speed and accuracy of measurements.



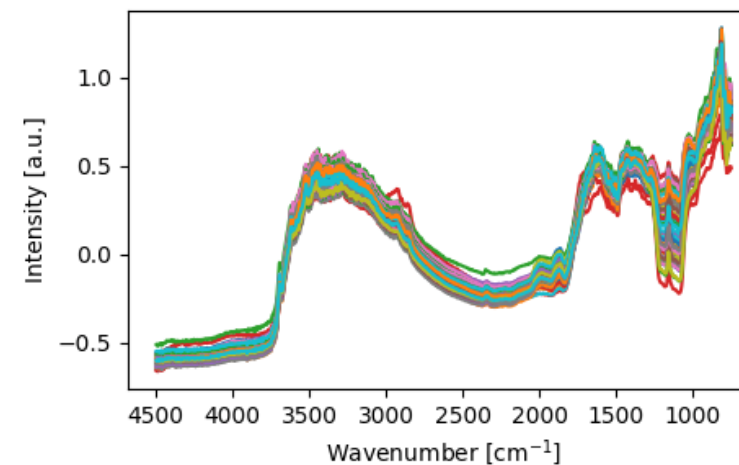
In-Field Operation

RemScan



Contaminated Soil

Mid-Infrared Spectra



CONTAMINATED

Binary
Classifier

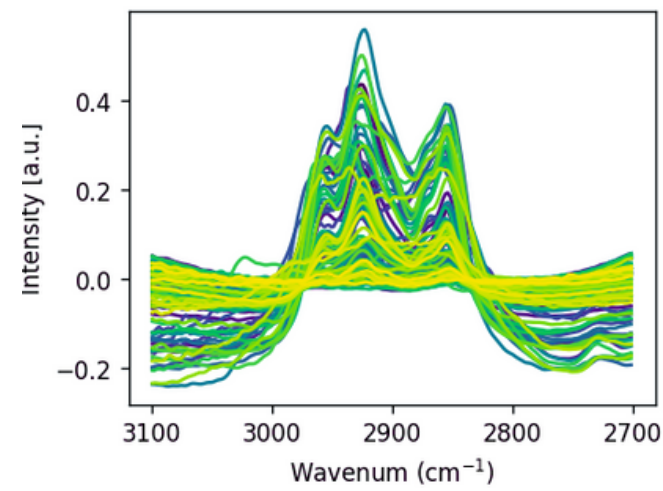


"CLEAN"

*Common Industry threshold = 1000 ppm/ 0.1% contamination

Methodology

TRAINING SPECTRA



LABELS

	filename	tpb	class_id
0	7832_13_01-29Jan14-1577_0000 (2)	100000	0
1	7832_13_01-29Jan14-1577_0001	100000	0
2	7832_13_01-29Jan14-1577_0002	100000	0
3	7832_13_01-29Jan14-1577_0003	100000	0
4	7832_13_01-29Jan14-1577_0004	100000	0
...	...	...	...
995	17738_12_09-03Feb14-1577_0001	1000	1
996	17738_12_09-03Feb14-1577_0002	1000	1
997	17738_12_09-03Feb14-1577_0003	1000	1
998	17738_12_09-03Feb14-1577_0004	1000	1
999	17738_12_10-03Feb14-1577_0000	781	1

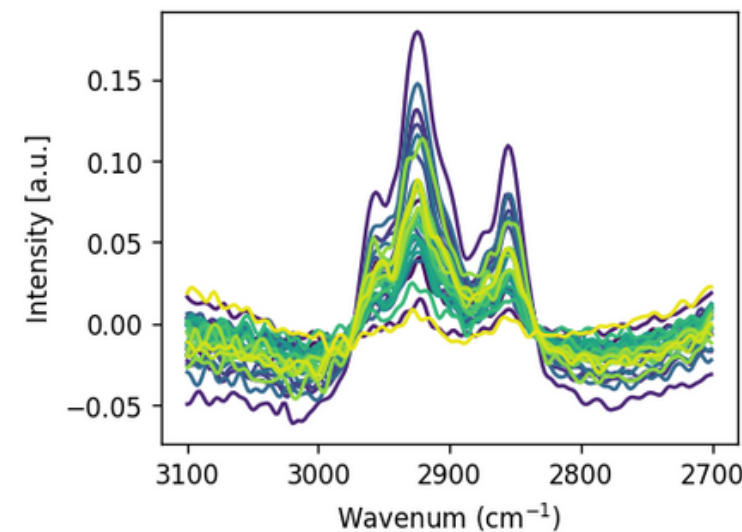
TRAINING
MODEL



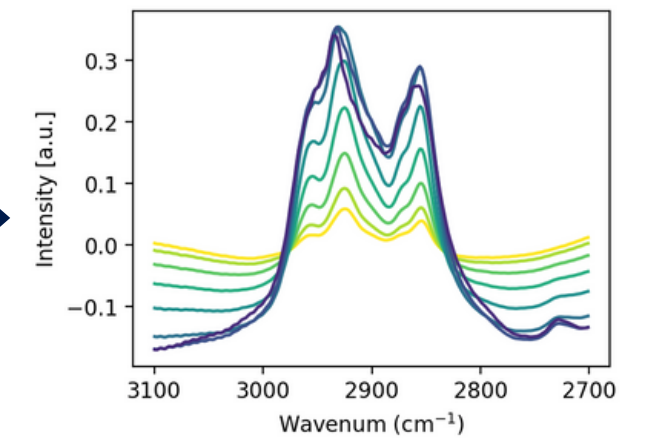
PREDICTION



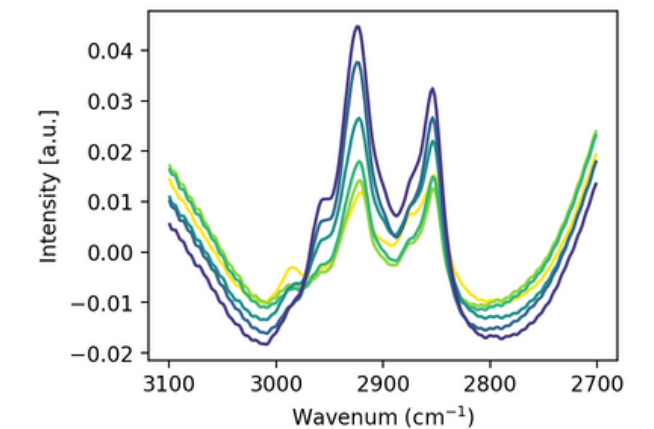
TEST SPECTRA



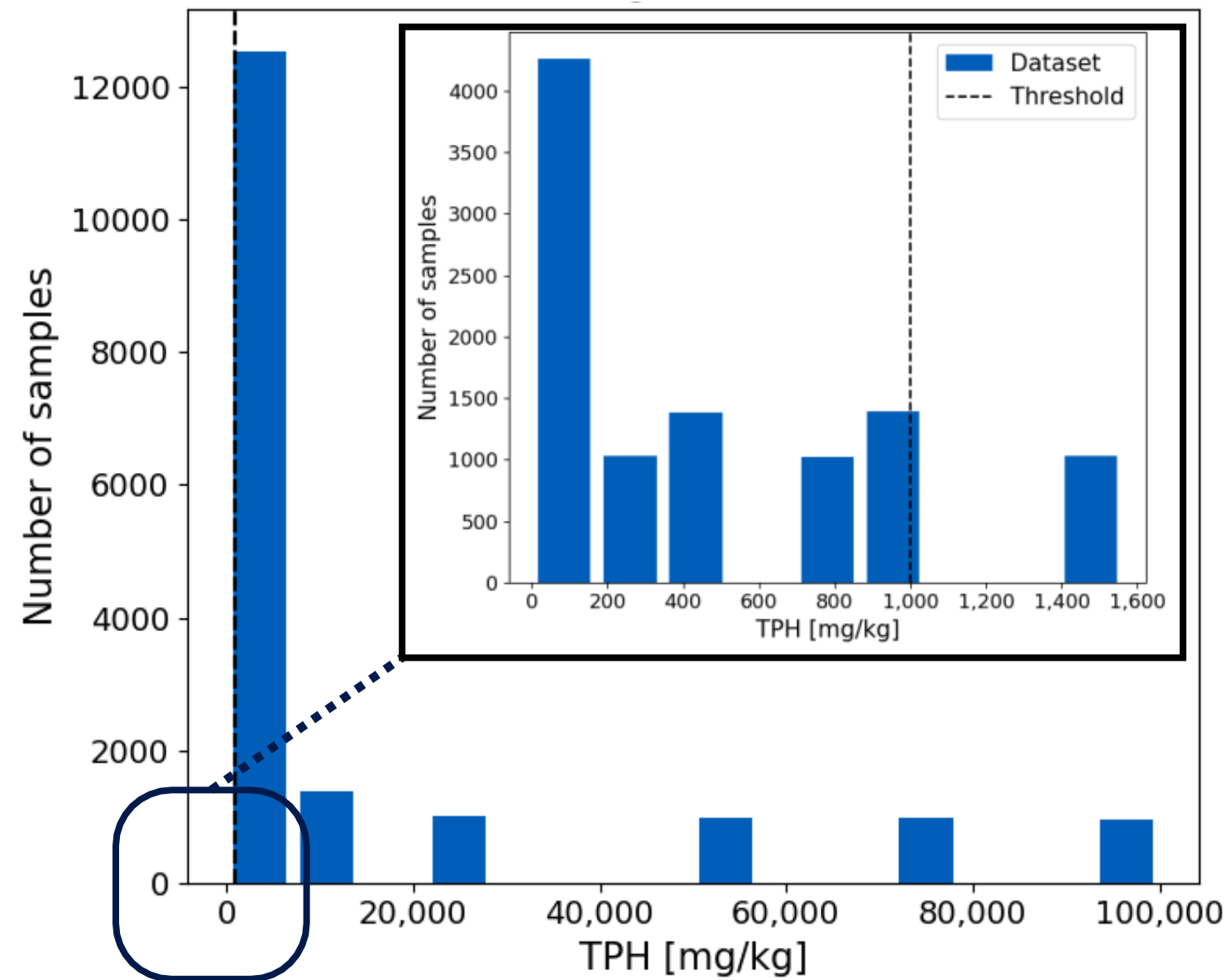
Contaminated:
TPH > 1000 ppm



Clean:
TPH < 1000 ppm



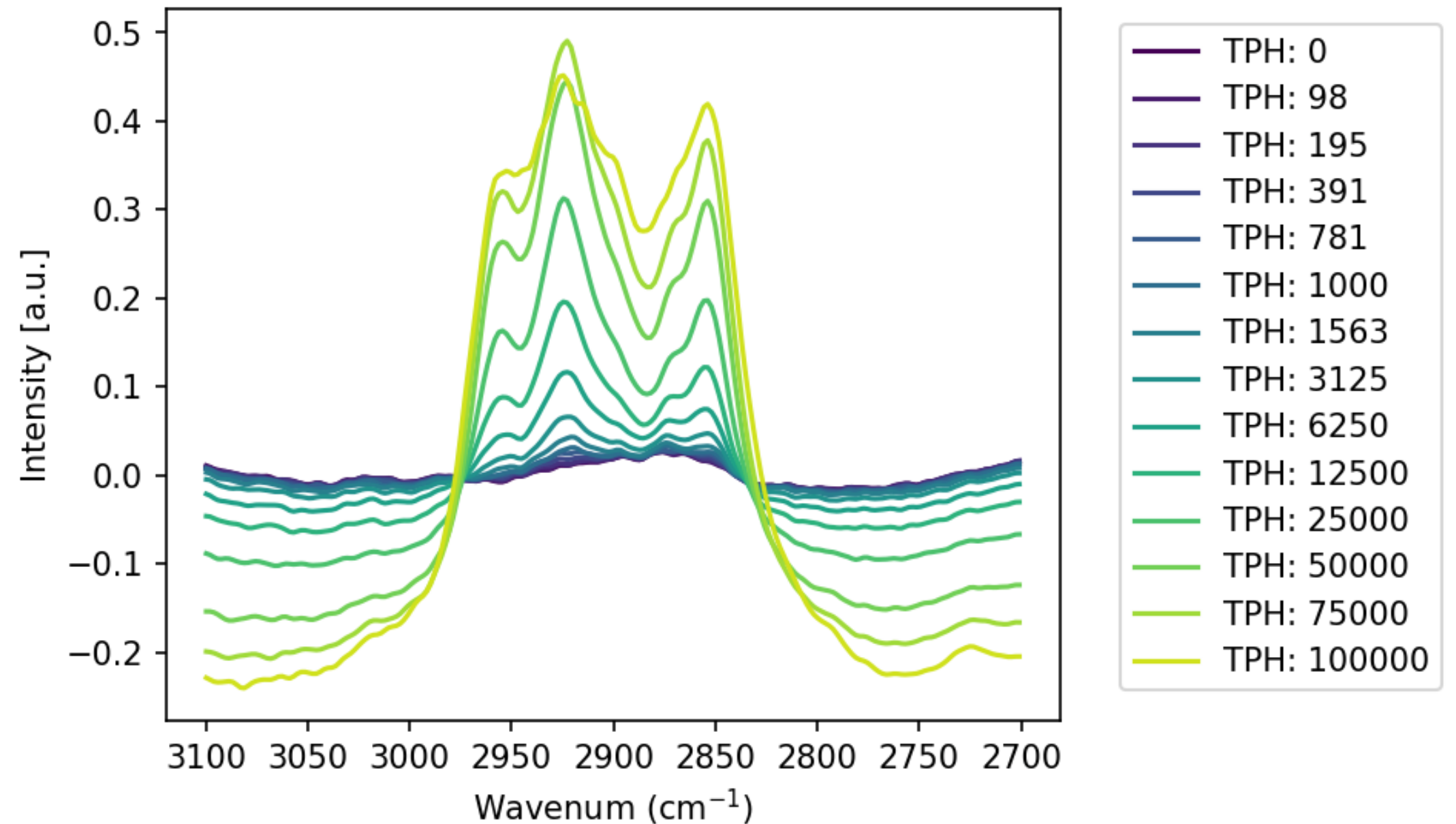
Training Dataset



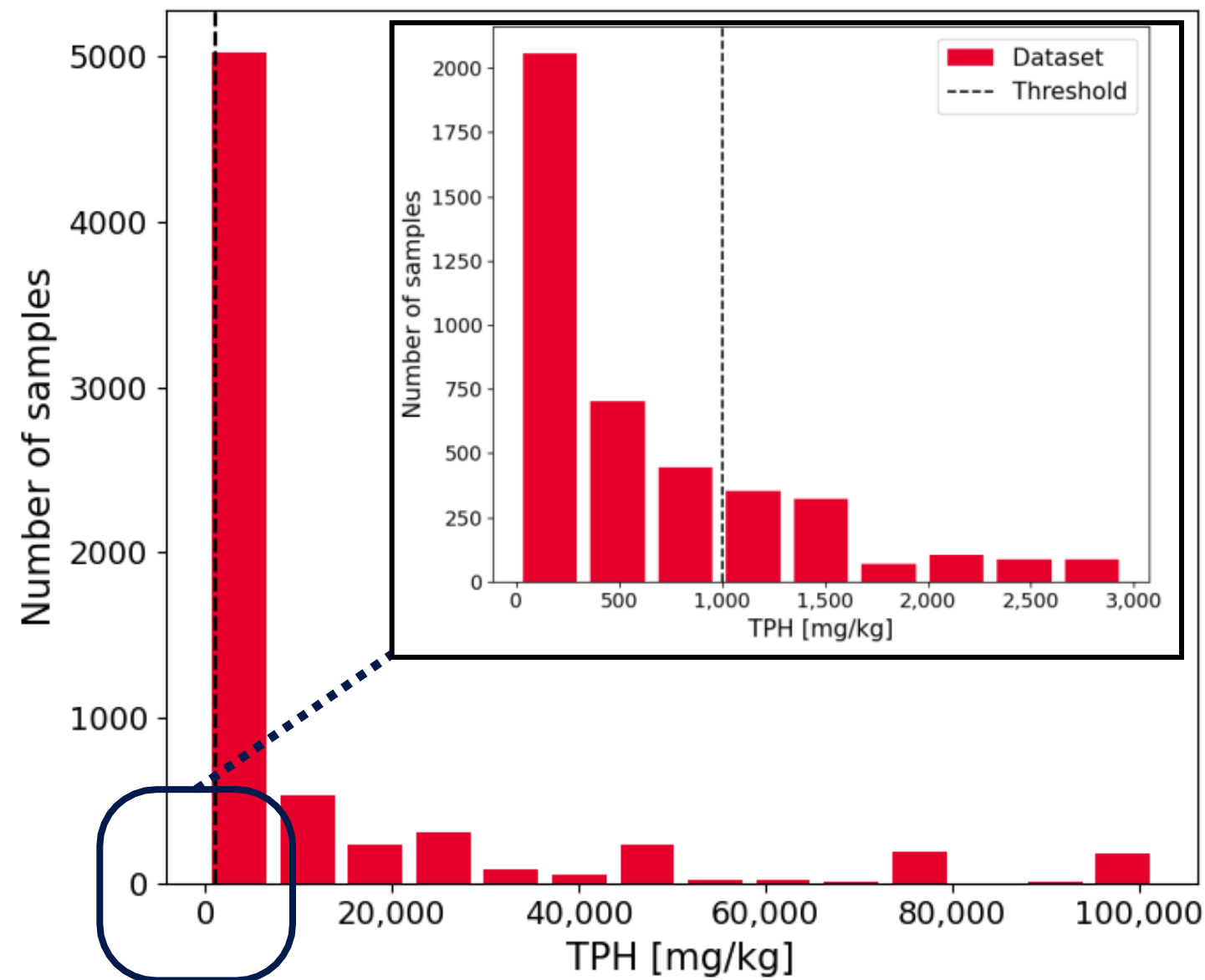
Total: 17,836

Samples >1000ppm = 8,750

Samples <1000ppm = 9,086



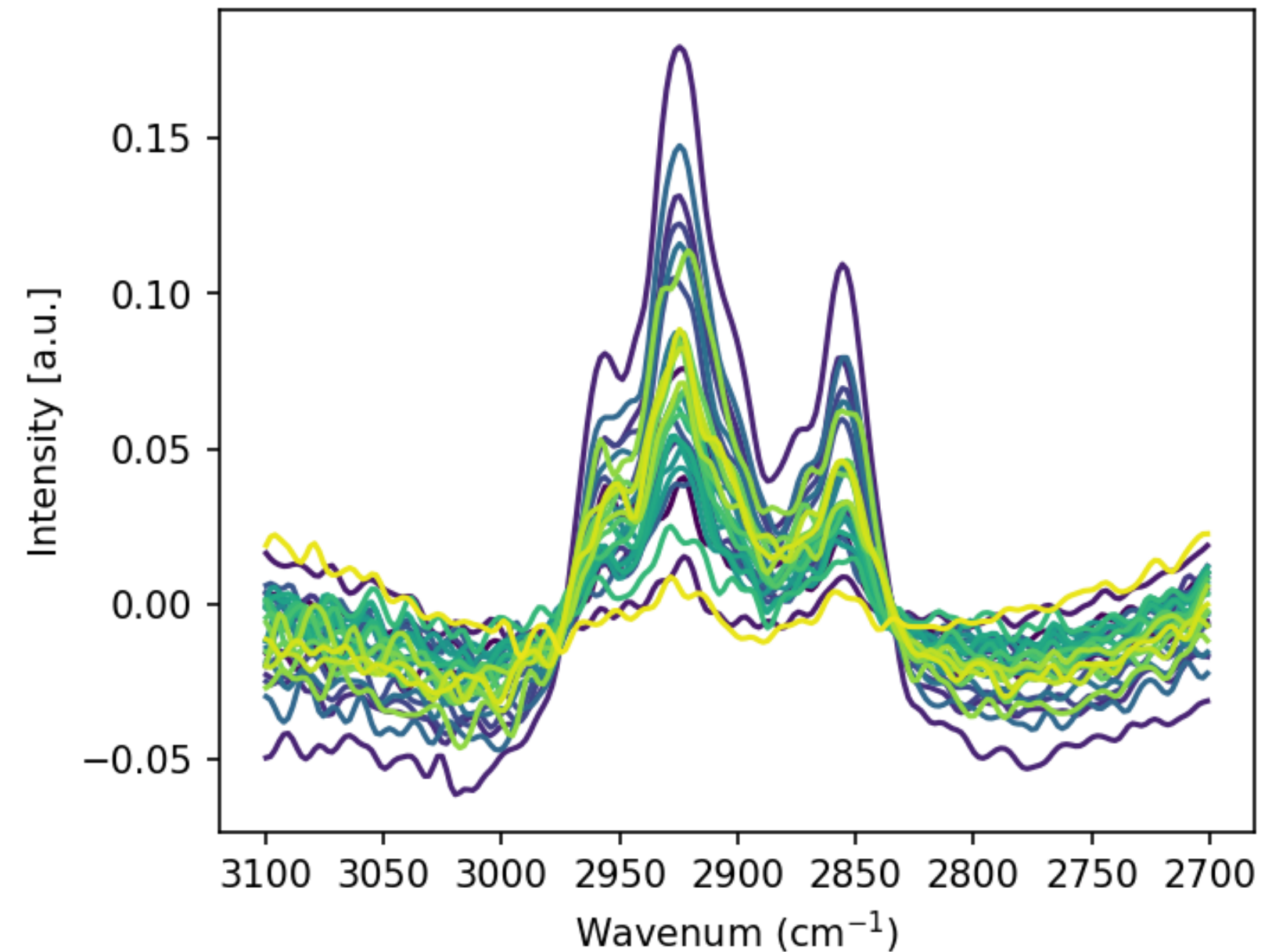
Testing Dataset



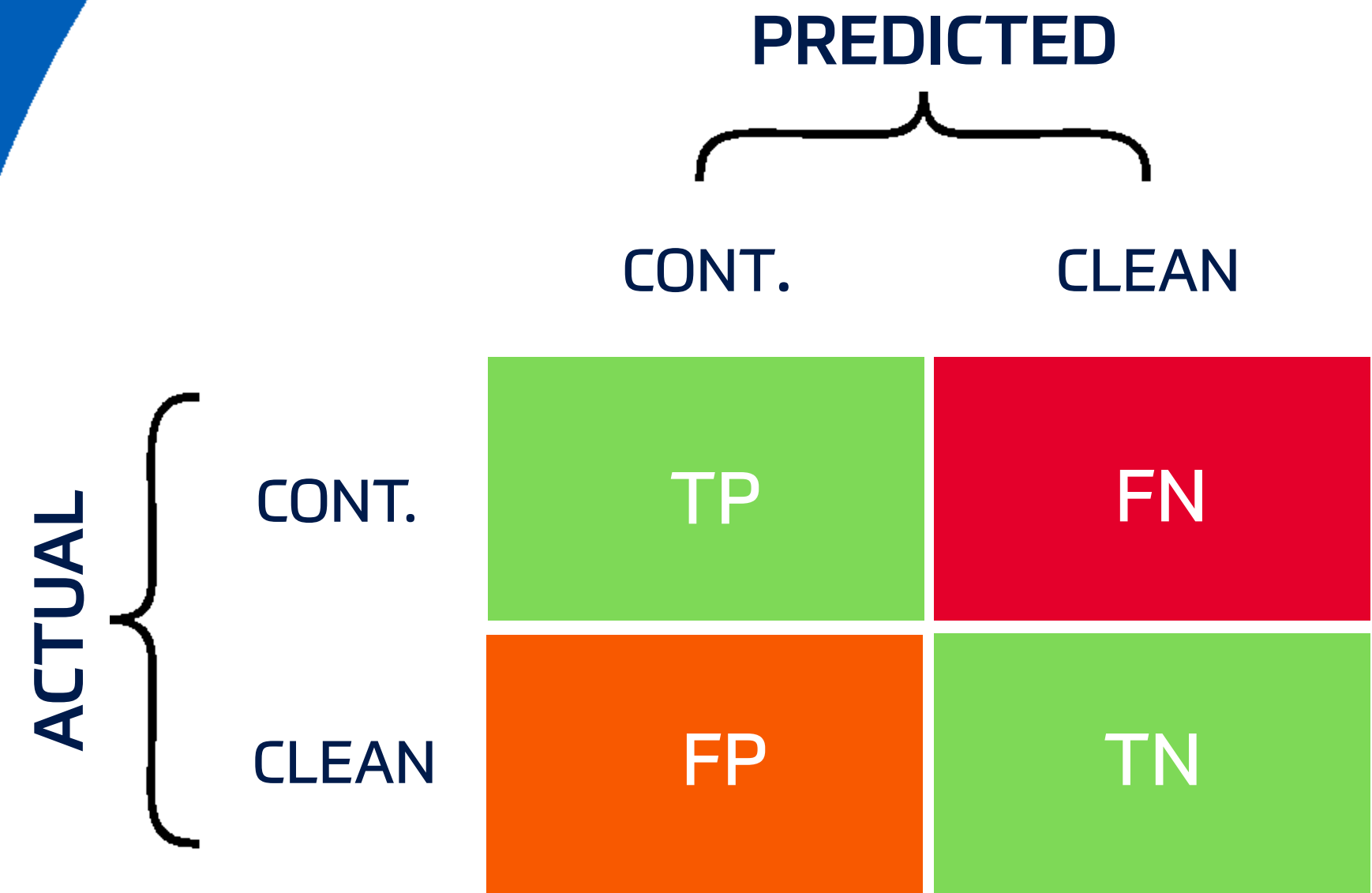
Total: 6,880

Samples >1000ppm = 3,485

Samples <1000ppm = 3,398



Confusion Matrix



CONT. = Contaminated

TP = True Positive (Cont.)

FP = False Positive

TN = True Negative (Clean)

FN = False Negative

Model refinement

Performance Optimisation

- Used 7 different classifiers
- Three metrics for assessing model performance
 - Accuracy
 - F1 score
 - Matthew's correlation coefficient (MCC)
- Get similar performance across all metrics.

Logistic regression

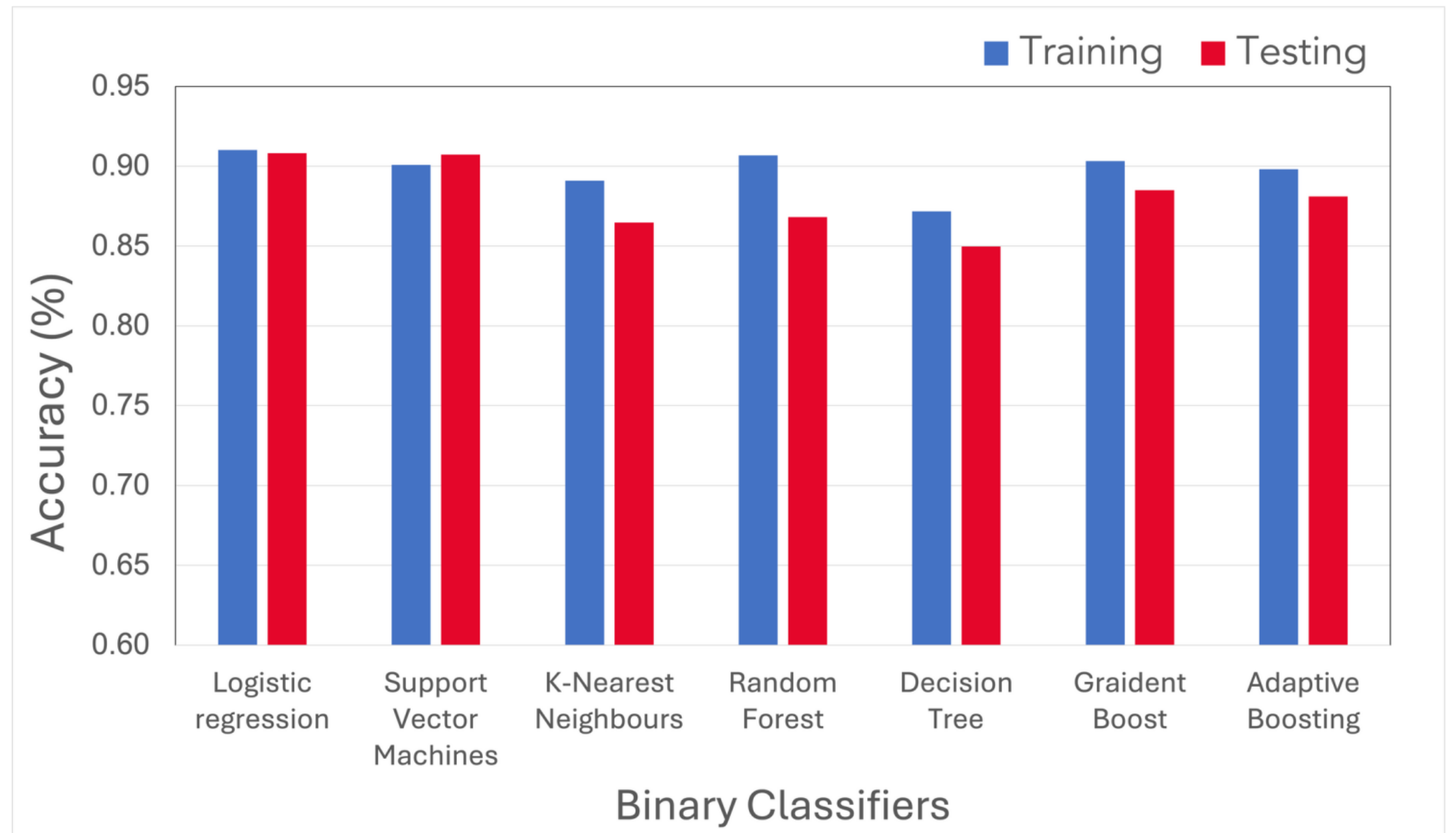
Training = 90.5 %

Testing = 90.8 %

Random forest

Training = 90.7 %

Testing = 86.8 %



Model refinement

Data Preparation

PREPROCESSING

- Tested three scenarios
- Best performance: Detrending

FEATURE SELECTION

- Select spectral regions of interest
- Three different combinations tested
- Best performance: Three regions

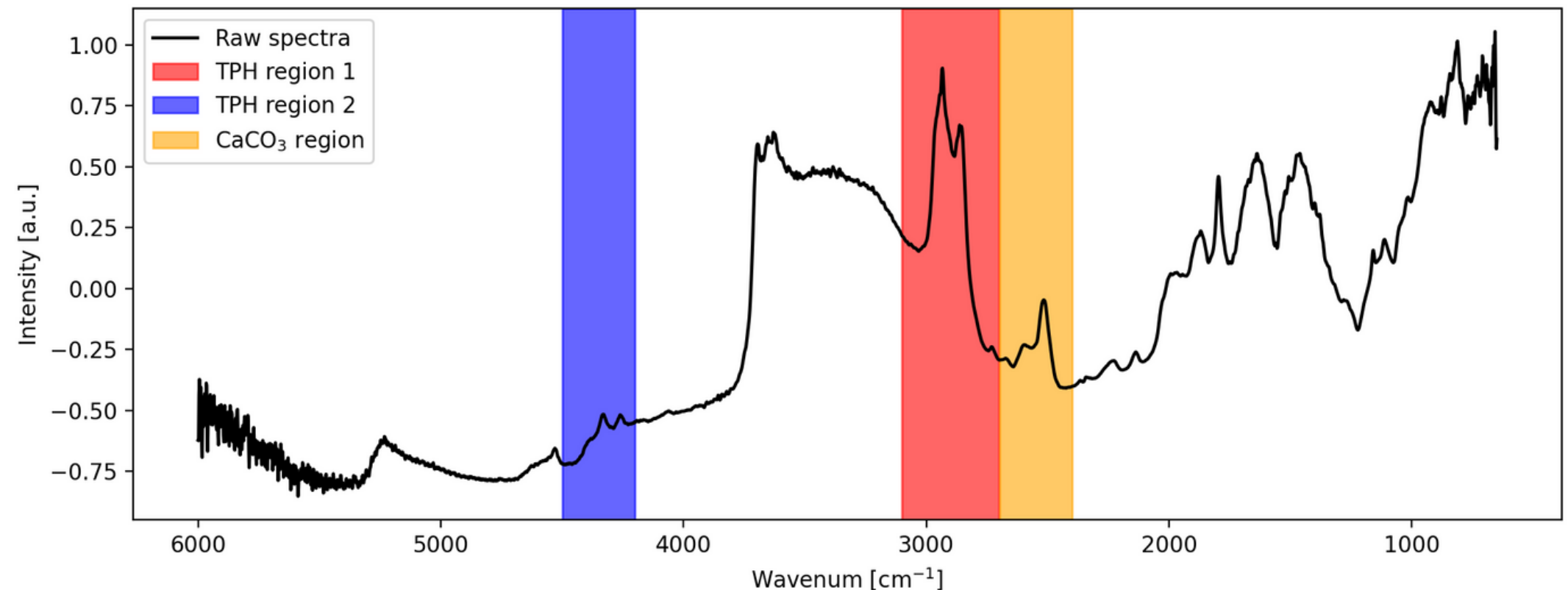


FIG: A typical mid-infrared spectrum of a contaminated sample. The red, blue and orange regions indicate the first TPH region, second TPH region and the calcium carbonate fingerprint region, respectively.

Interfering signals

Calcium Carbonate [CaCO_3]

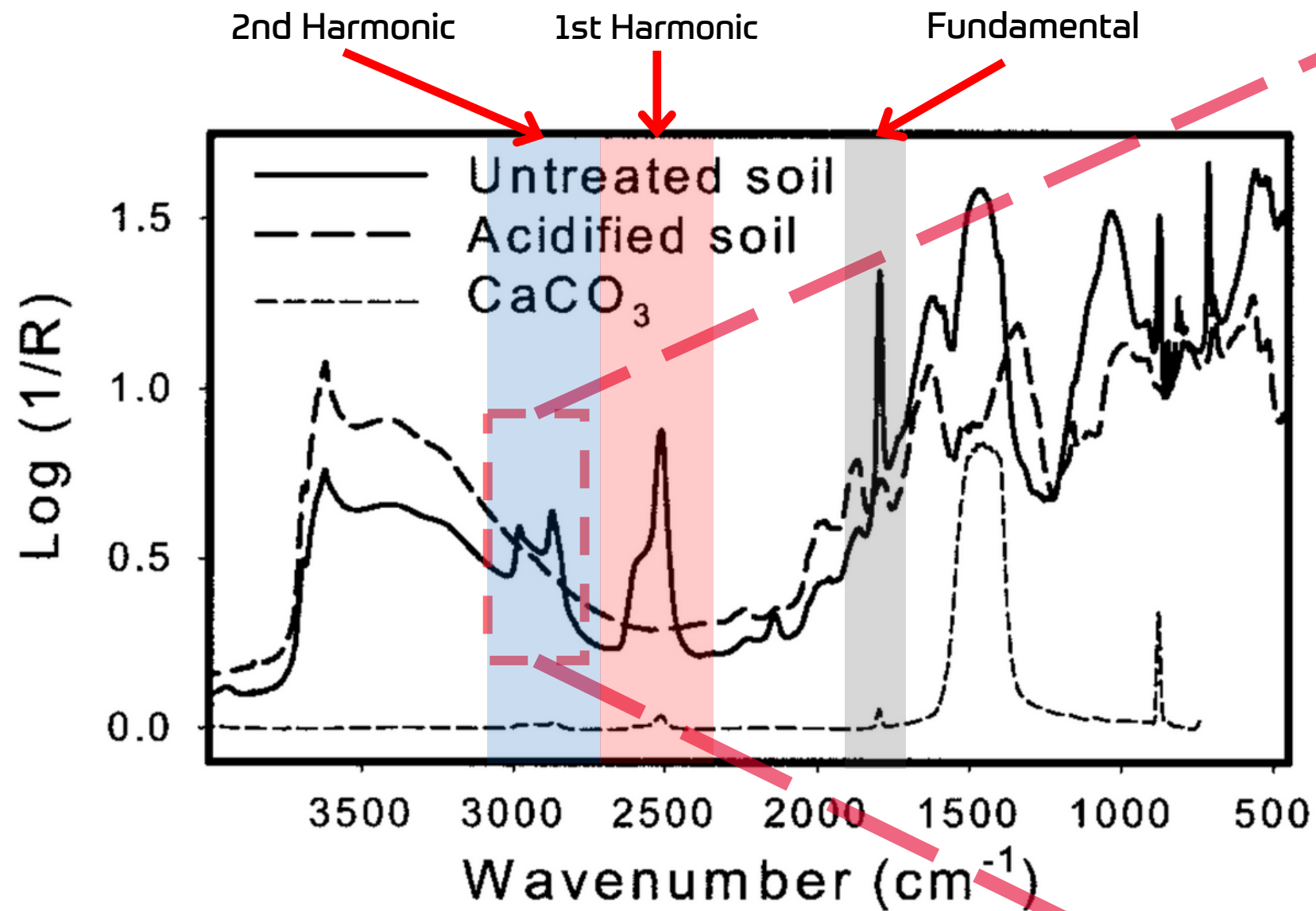
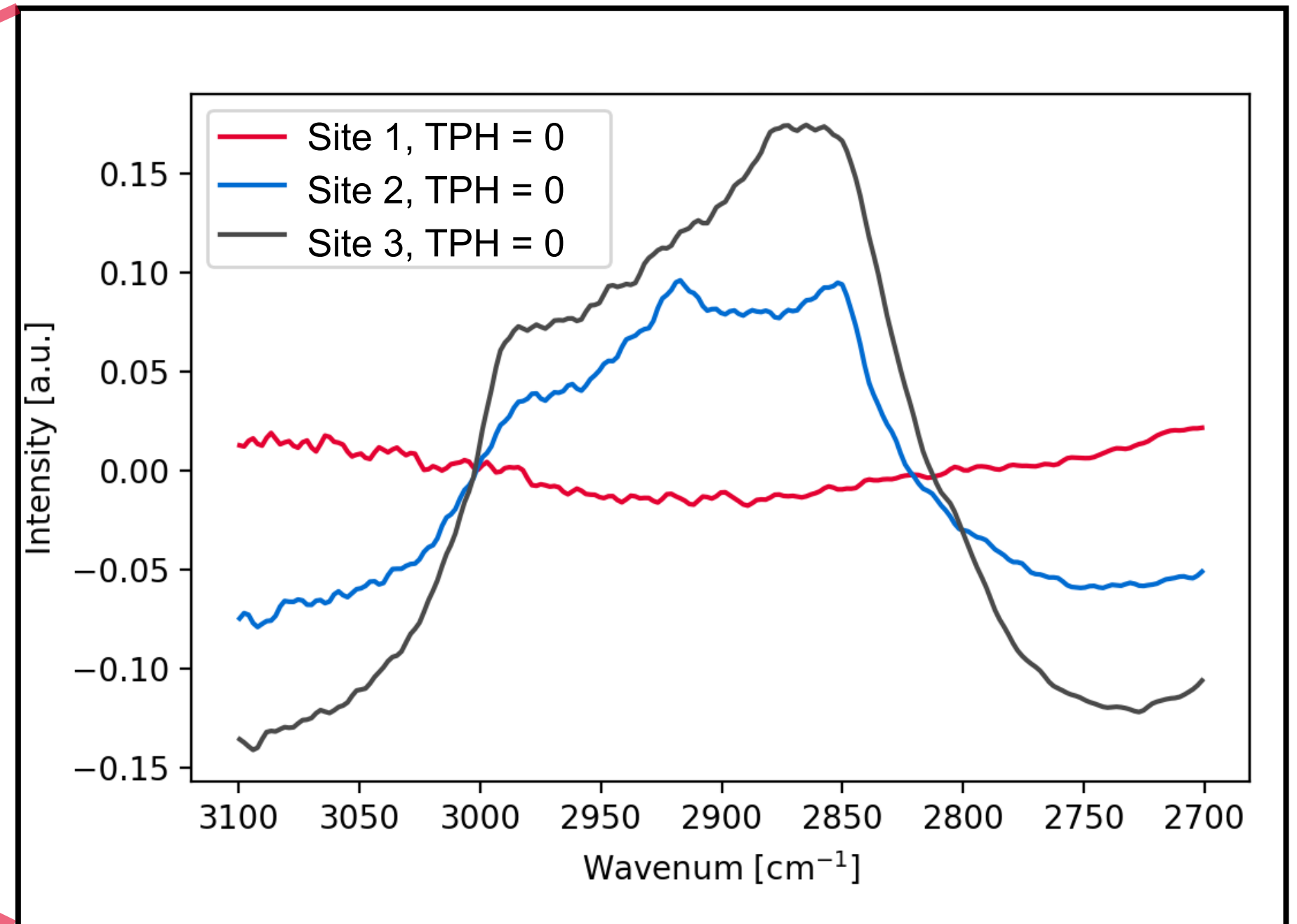


FIG: Comparison of mid-infrared and near-infrared spectra of a highly calcareous soil before and after treatment with acid for removal of carbonates. The carbonate (i.e., CaCO_3) spectrum is included for additional comparison.

Reference: G. McCarty et al., Mid-Infrared and Near-Infrared Diffuse reflectance spectroscopy for Soil Carbon Measurements.

DOI: [10.2136/sssaj2002.6400](https://doi.org/10.2136/sssaj2002.6400)



Interfering signals

Soil Organic Carbon (SOC)

- Known overlap between TPH-sensitive IR peak & natural organic matter

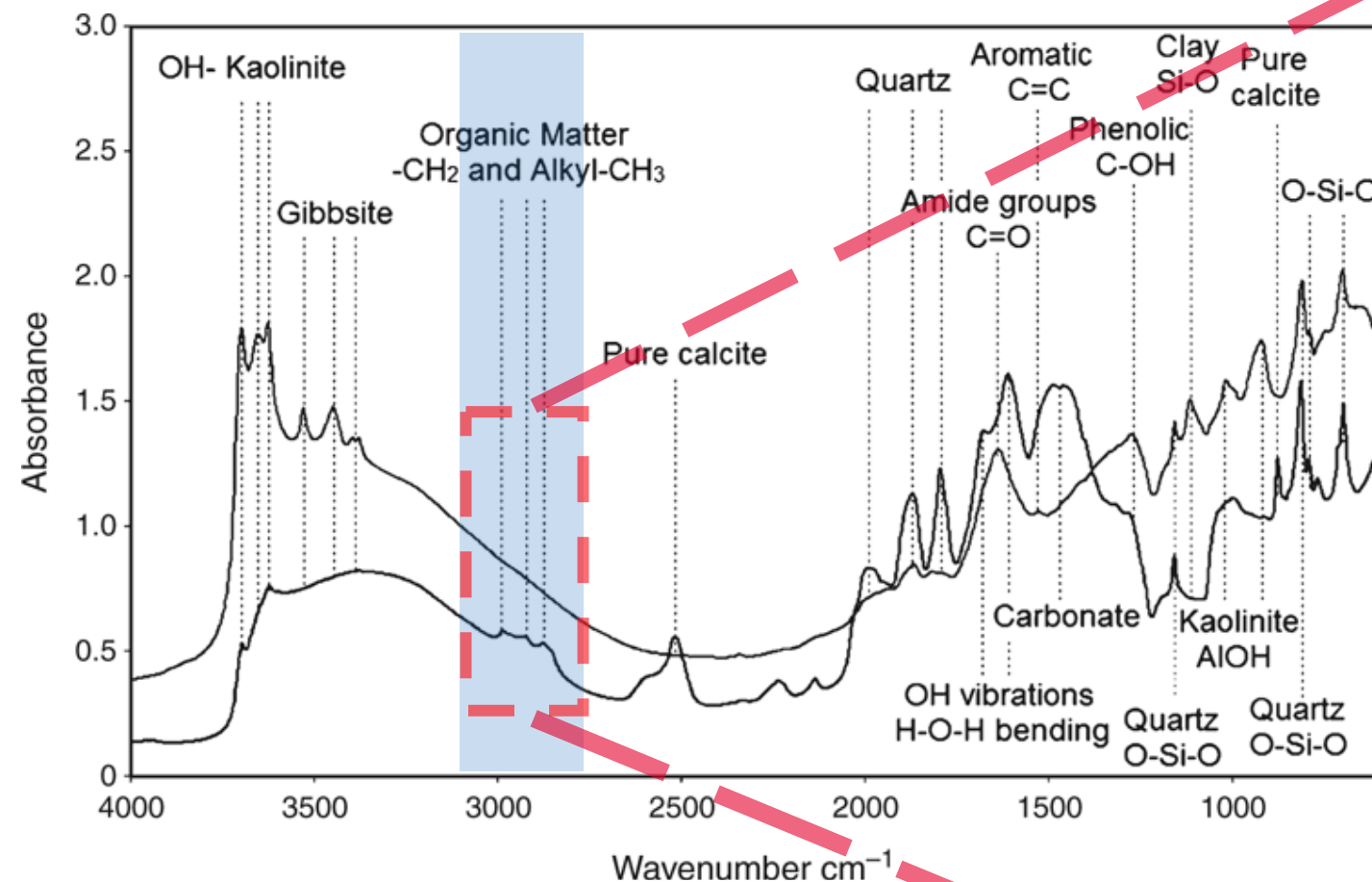
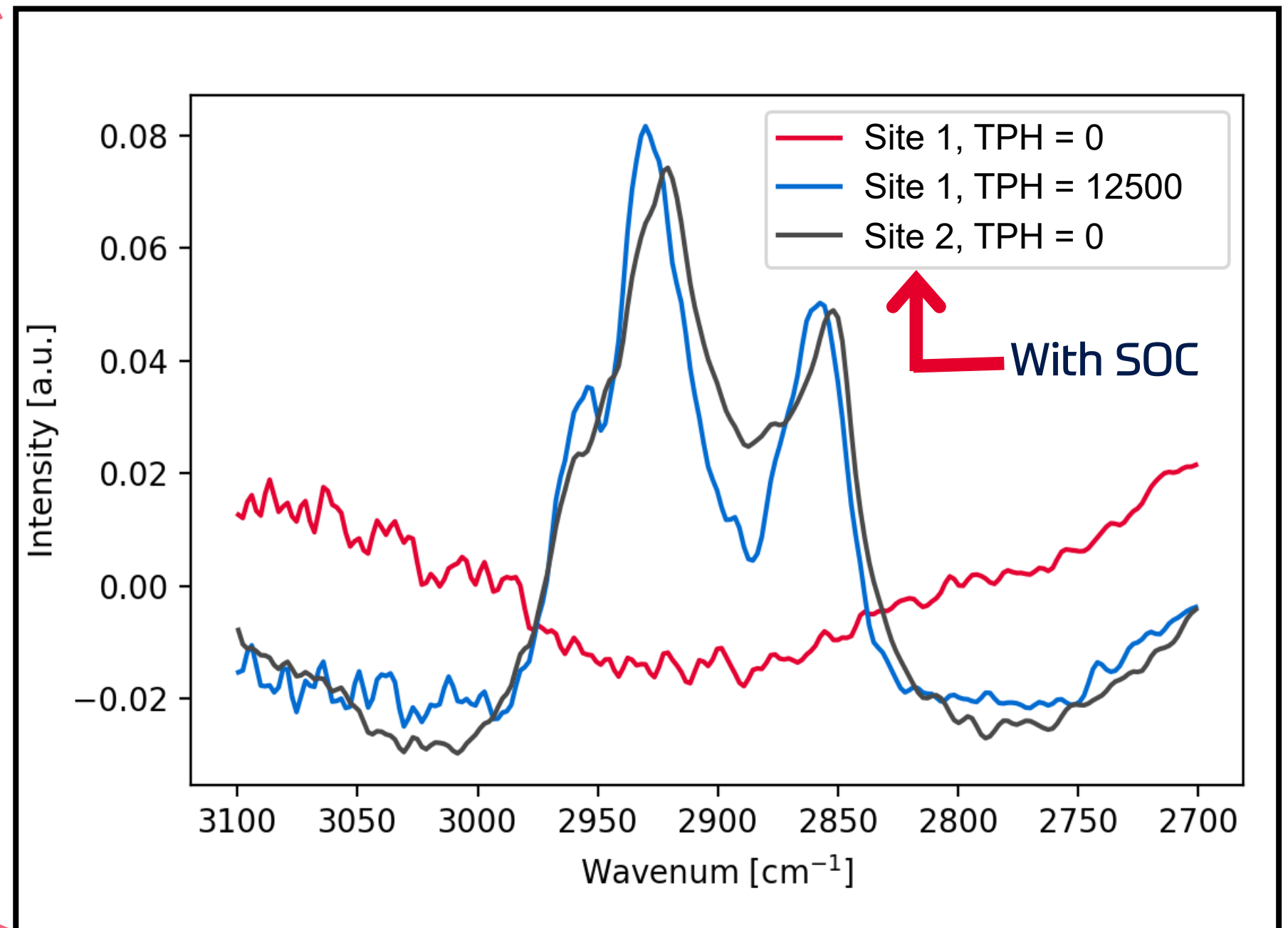


FIG: Representative soil mid-IR spectrum showing absorptions related to the mineral and organic composition of soil.

Reference: F. Le Guillou et al., How does grinding affect the mid-infrared spectra of soil and their multivariate calibrations to texture and organic carbon?.

DOI: 10.1071/SR15019

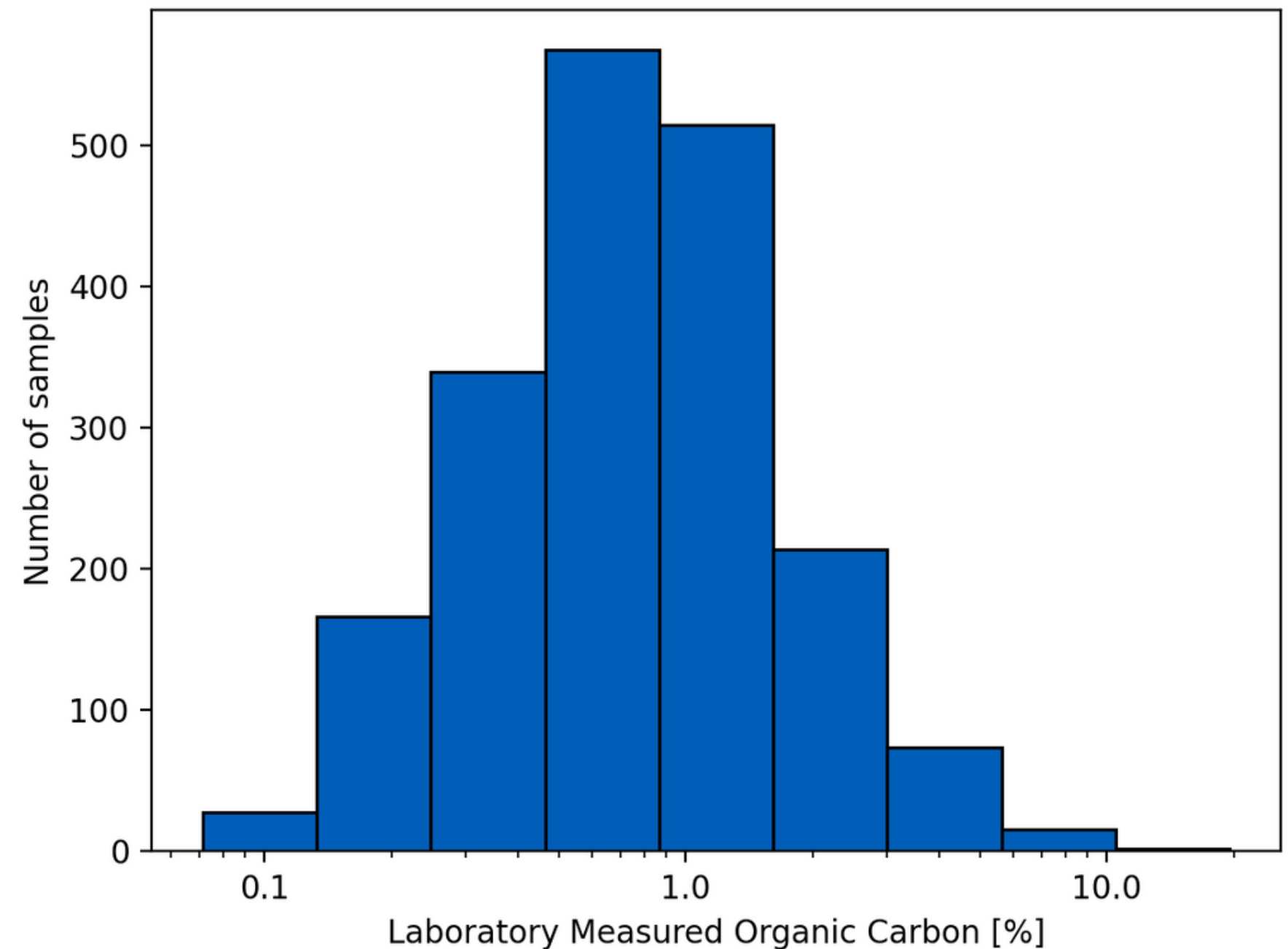


Interfering signals

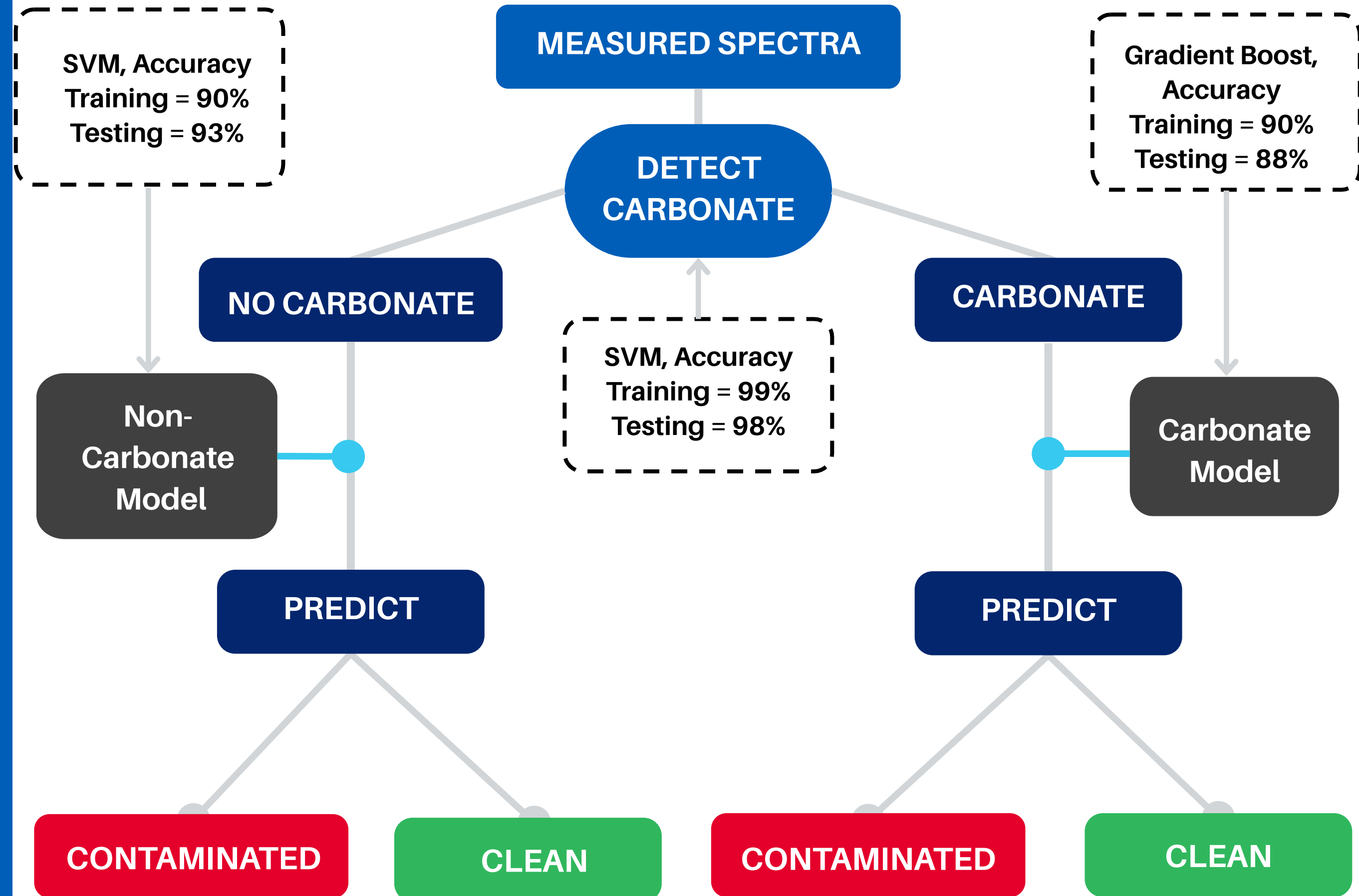
Soil Organic Carbon (SOC)

- Clean samples collected across Australia
- SOC content measured by accredited laboratory [Dry Combustion]
- Classifier: Support Vector Machines (SVM)
 - Accuracy: 99%
 - Confusion matrix

		Predictions	
		Cont.	Clean
Actual	Cont.	0	0
	Clean	2	1919



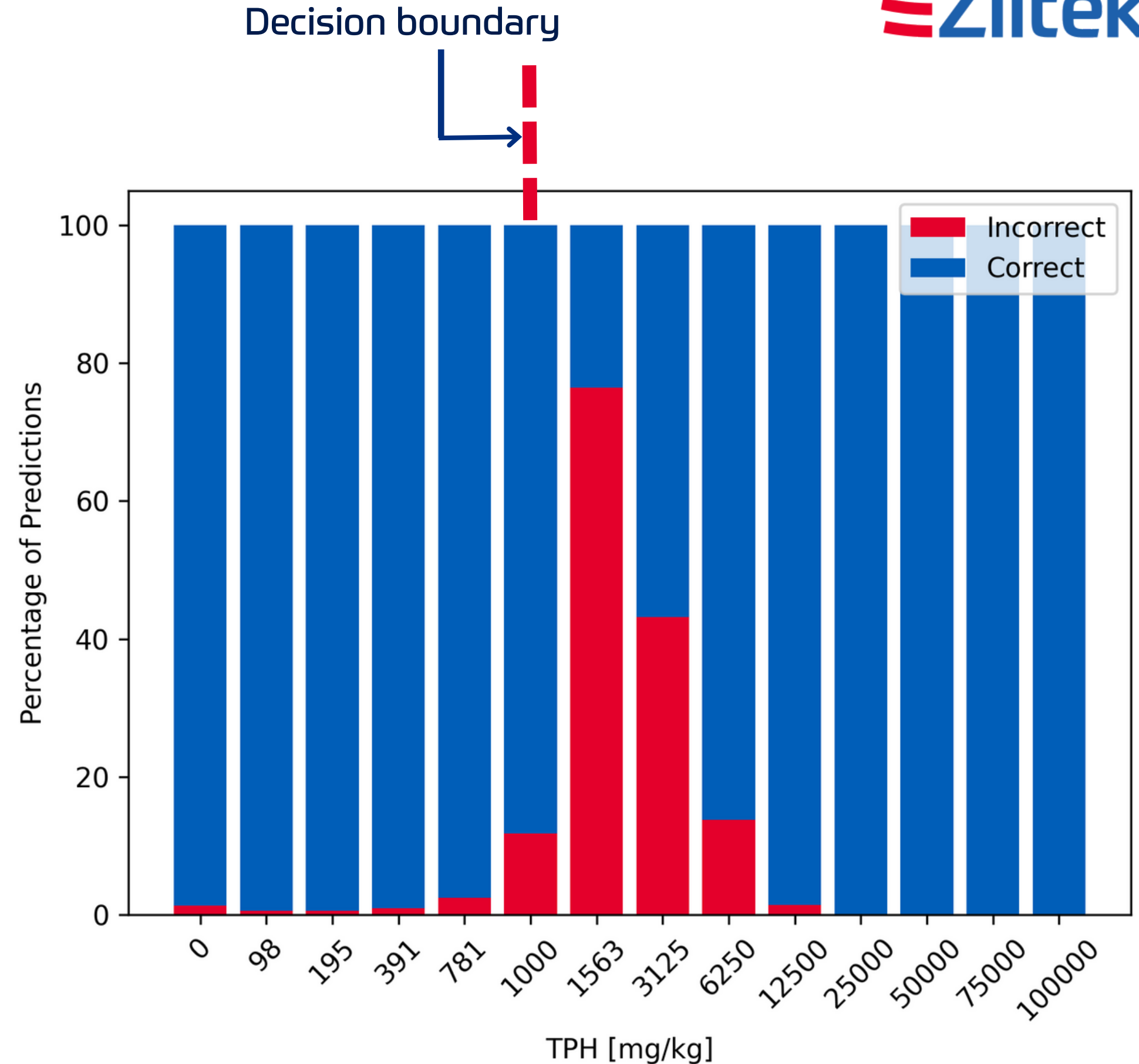
Model Workflow



Model Summary

NON-CARBONATE MODEL

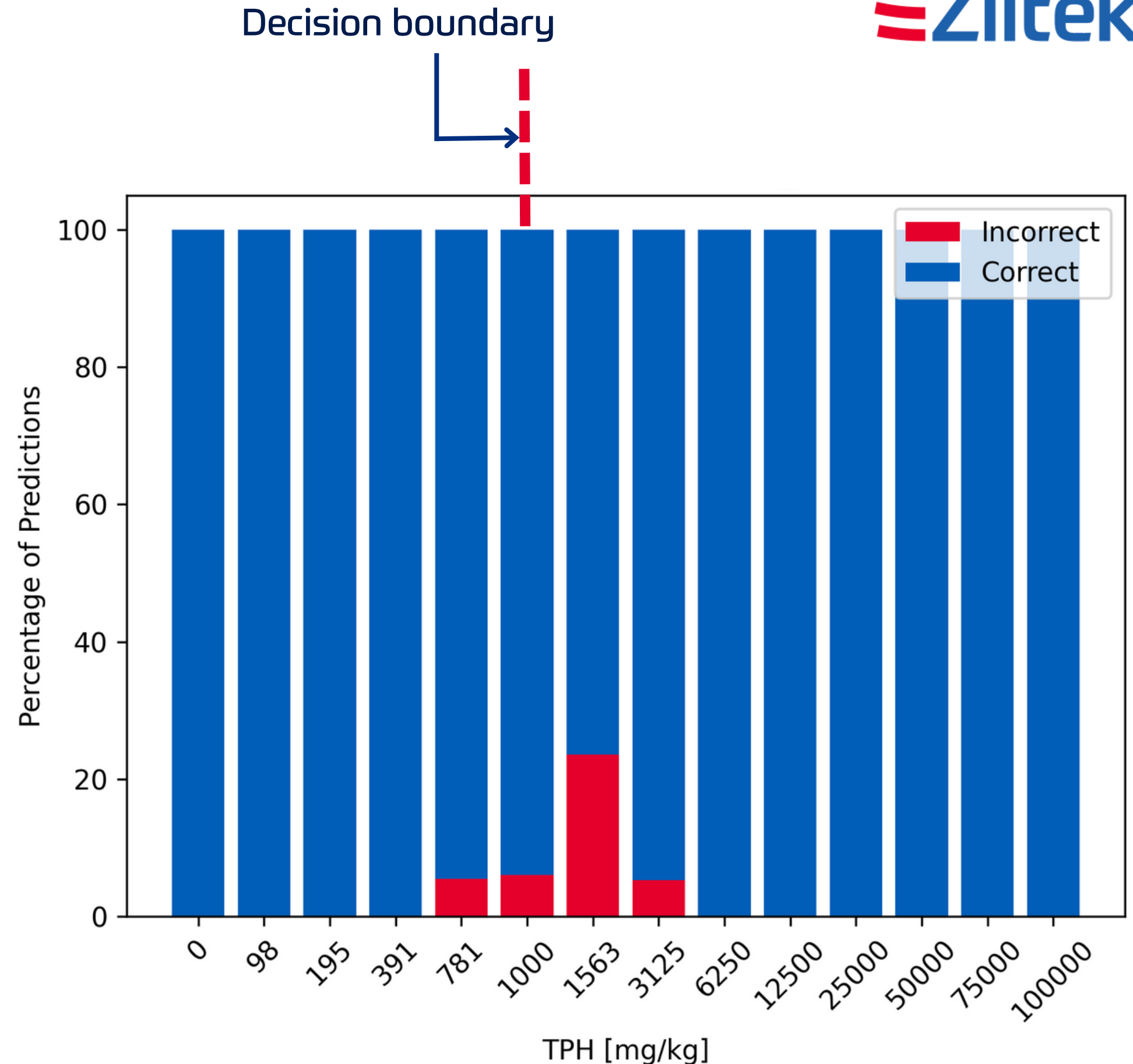
- Filtered dataset
 - Training = 13,984
 - Testing = 5,008
- Preprocessing = Detrend
- Best Classifier = SVM
- Accuracy
 - Training = 90%
 - Testing = 93%
- Highest misclassified samples just above the threshold
- Misclassified samples at lower TPH values likely due to SOC.



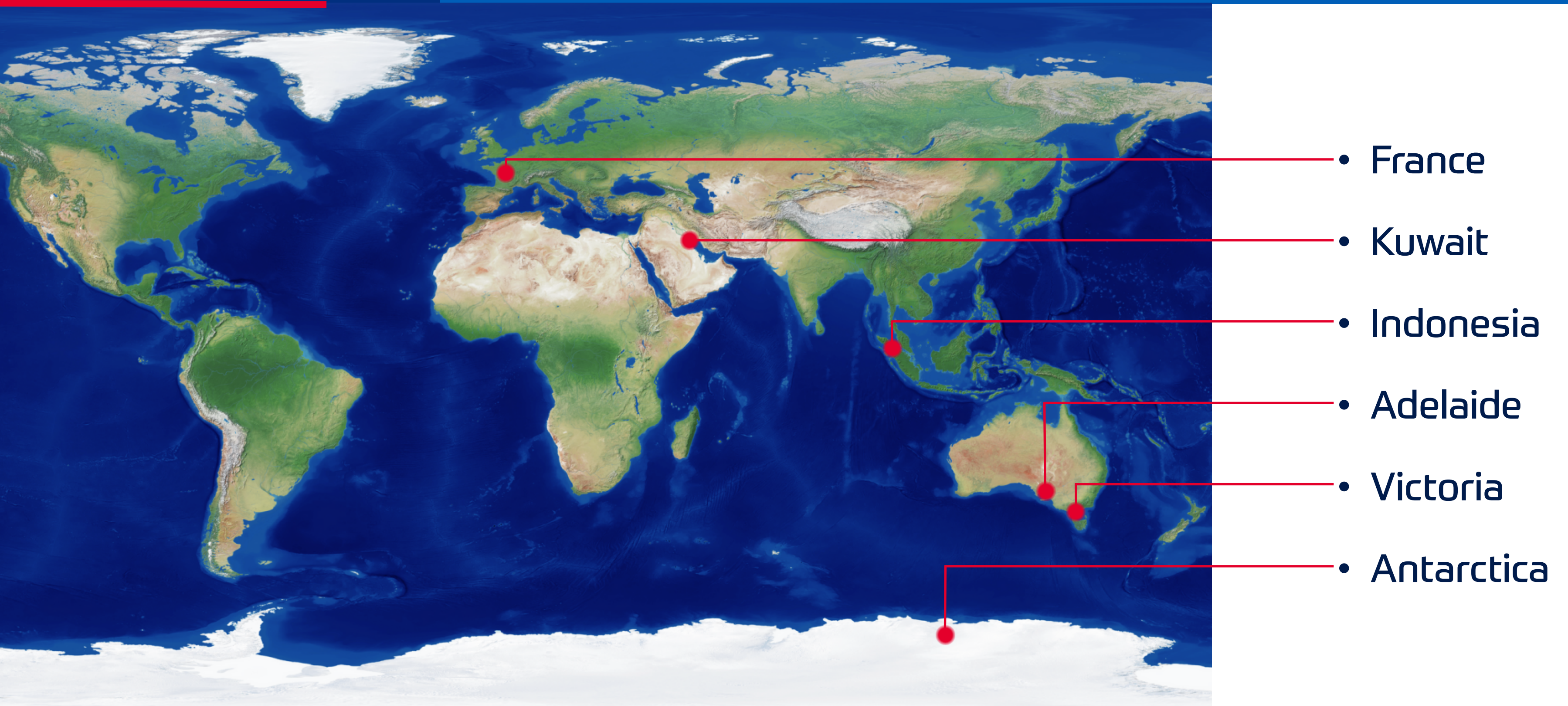
Model Summary

CARBONATE MODEL

- Model built exclusively on carbonate samples.
 - Training = 3,852
 - Testing = 1,872
- Preprocessing: Detrend
- Best Classifier = Gradient Boost
- Accuracy
 - Training = 90.1%
 - Testing = 88.4%
- Require more samples for future optimisation.



Test Results

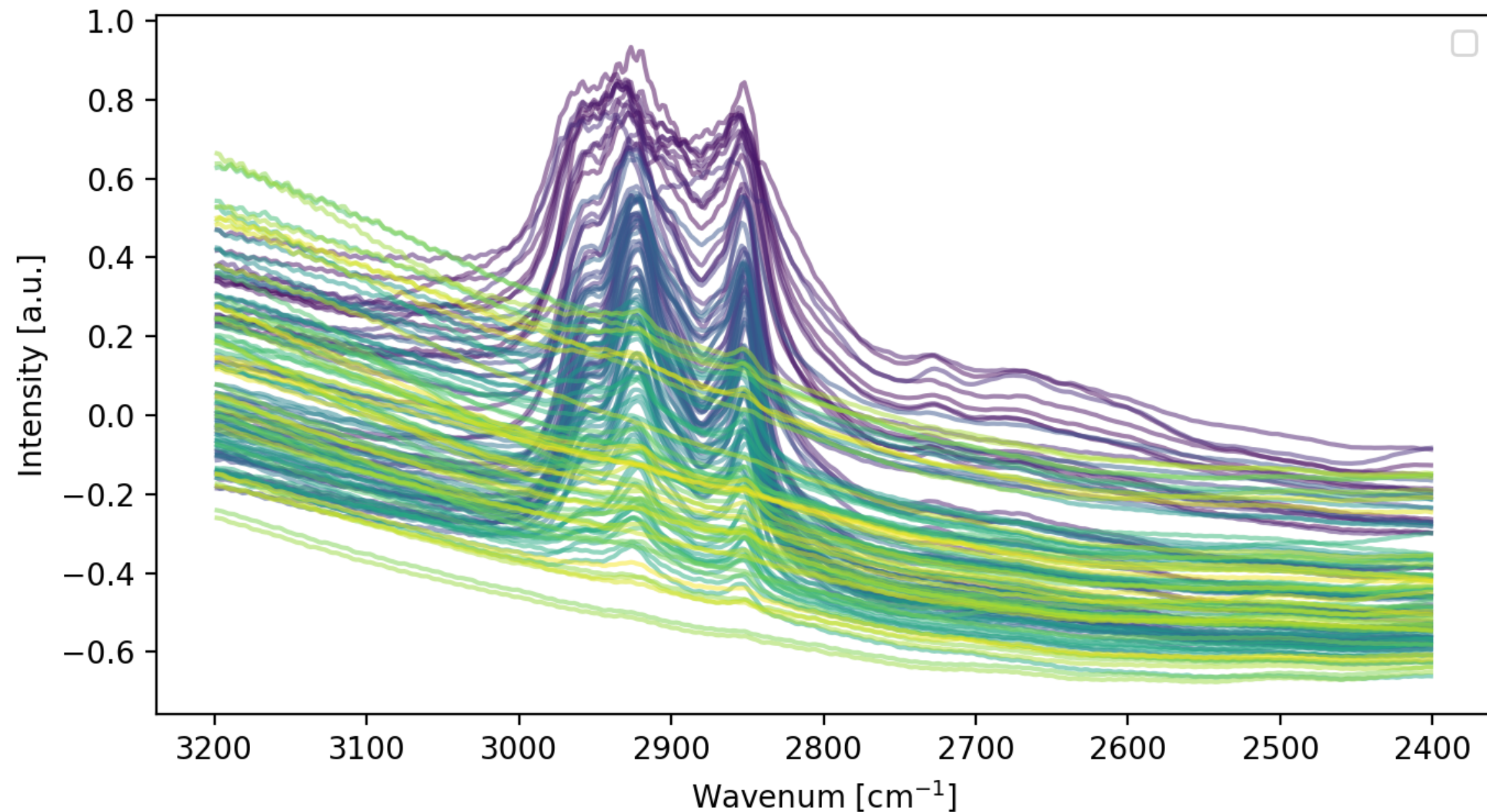


Results

TEST CASE 1 - Indonesia

CONFUSION MATRIX $\begin{bmatrix} 1218 & 10 \\ 80 & 730 \end{bmatrix}$

- Project Details
 - Sumatran oil fields
 - Contaminant: Crude Oil
 - Large scale, multi-year project
- Model: Non-Carbonate
- Number of samples: 2038
 - Class A: 1218 (>1000)
 - Class B: 810 (<1000)
- Classifier: SVM
 - Accuracy: 95%
 - F1 Score: 0.94
 - MCC: 0.90

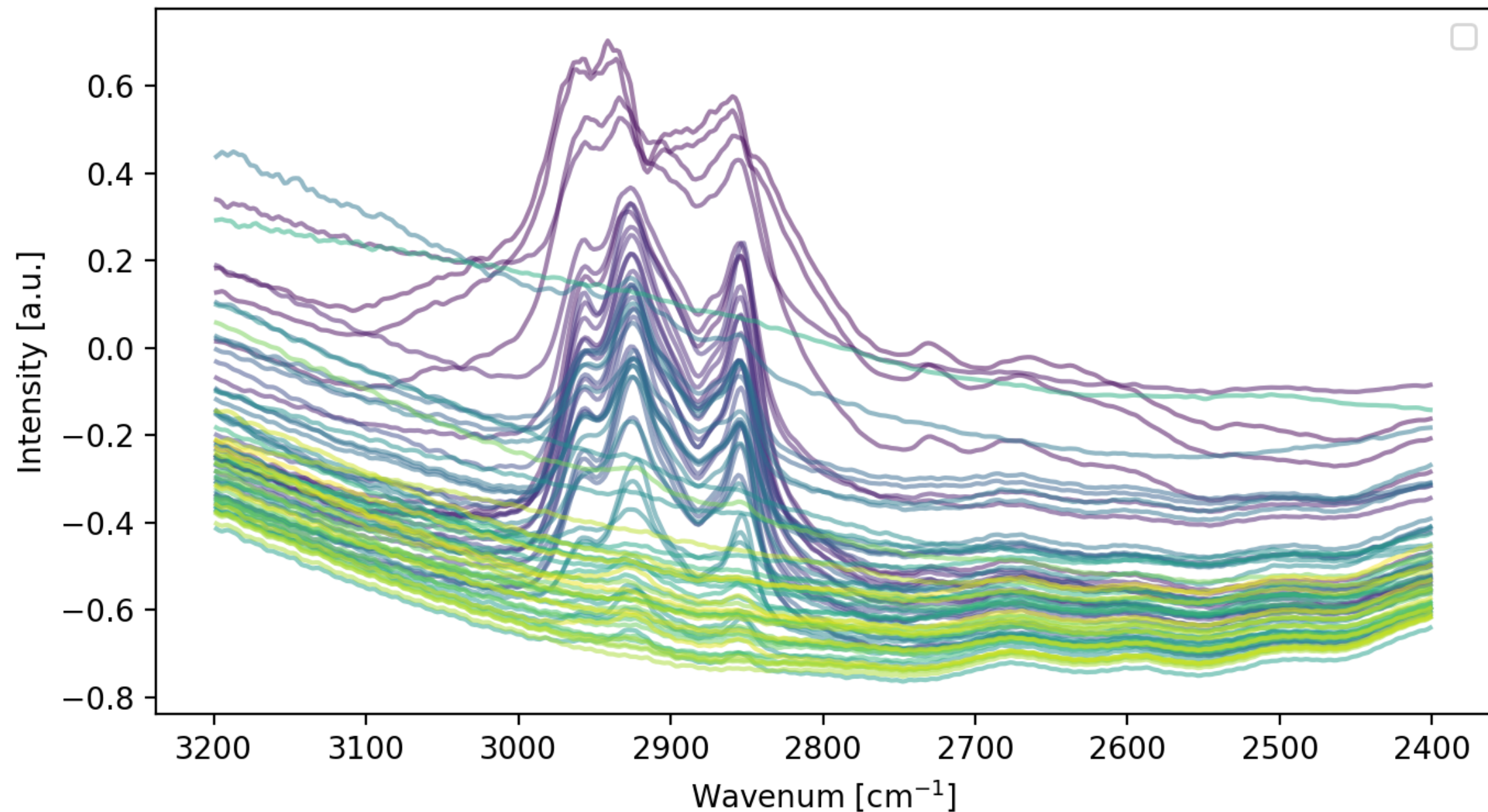


Results

TEST CASE 2 - Coastal Victoria

CONFUSION MATRIX $\begin{bmatrix} 117 & 0 \\ 0 & 361 \end{bmatrix}$

- Project Details
 - Coastal Wilderness
 - Contaminant: Crude Oil
 - Medium scale, 18 months project
- Model: Non-Carbonate
- Number of samples: 538
 - Class A: 361 (>1000)
 - Class B: 177 (<1000)
- Classifier: SVM
 - Accuracy: 100%
 - F1 Score: 1.0
 - MCC: 1.0

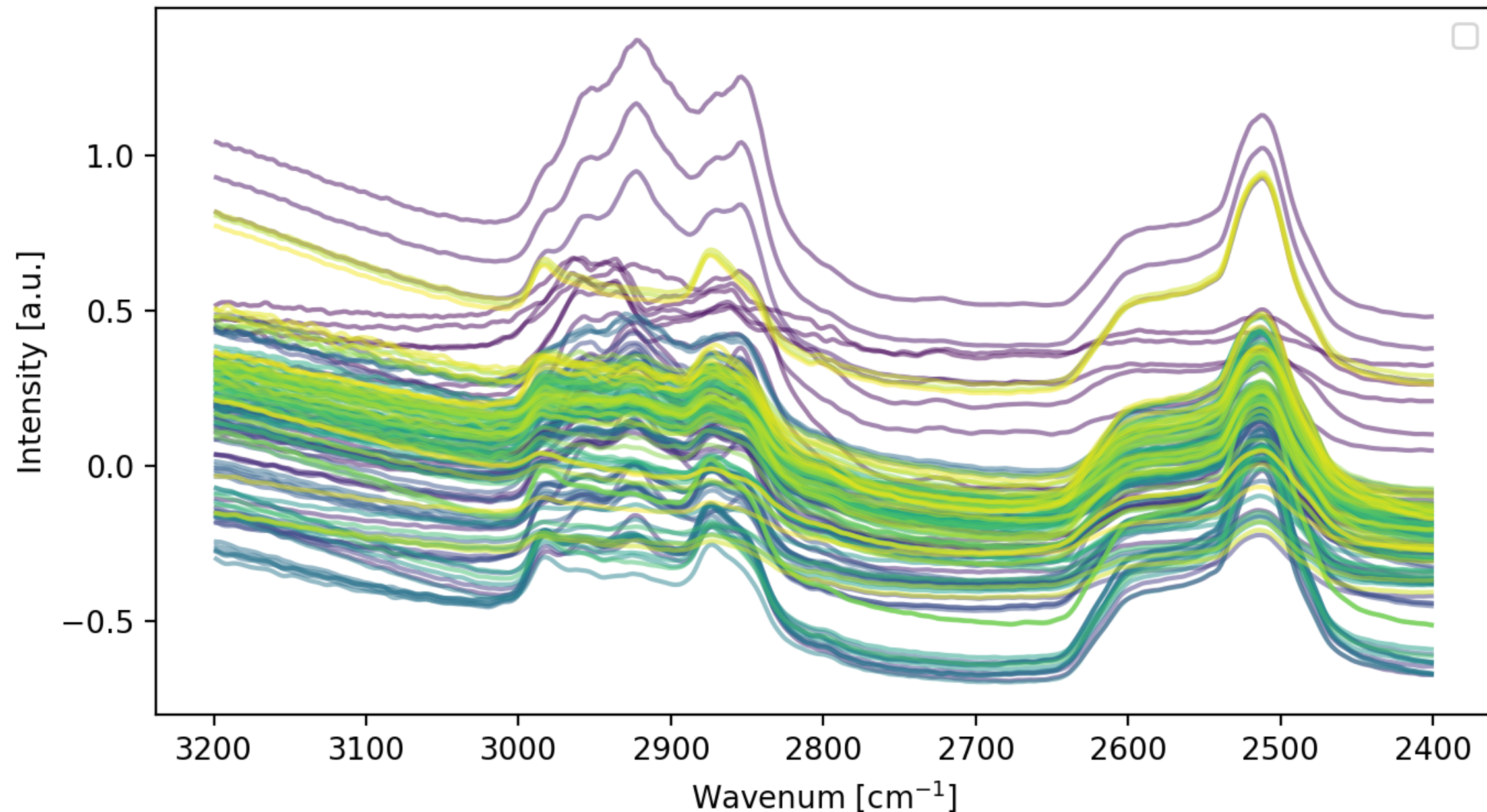


Results

TEST CASE 3 - France

CONFUSION MATRIX $\begin{bmatrix} 389 & 83 \\ 29 & 396 \end{bmatrix}$

- Project Details:
 - Industrial site
 - Contaminant: Diesel
 - Small scale, several weeks
- Model: Carbonate
- Number of samples: 897
 - Class A: 472 (>1000)
 - Class B: 452 (<1000)
- Classifier: Gradient Boost
 - Accuracy: 88%
 - F1 Score: 0.88
 - MCC: 0.76



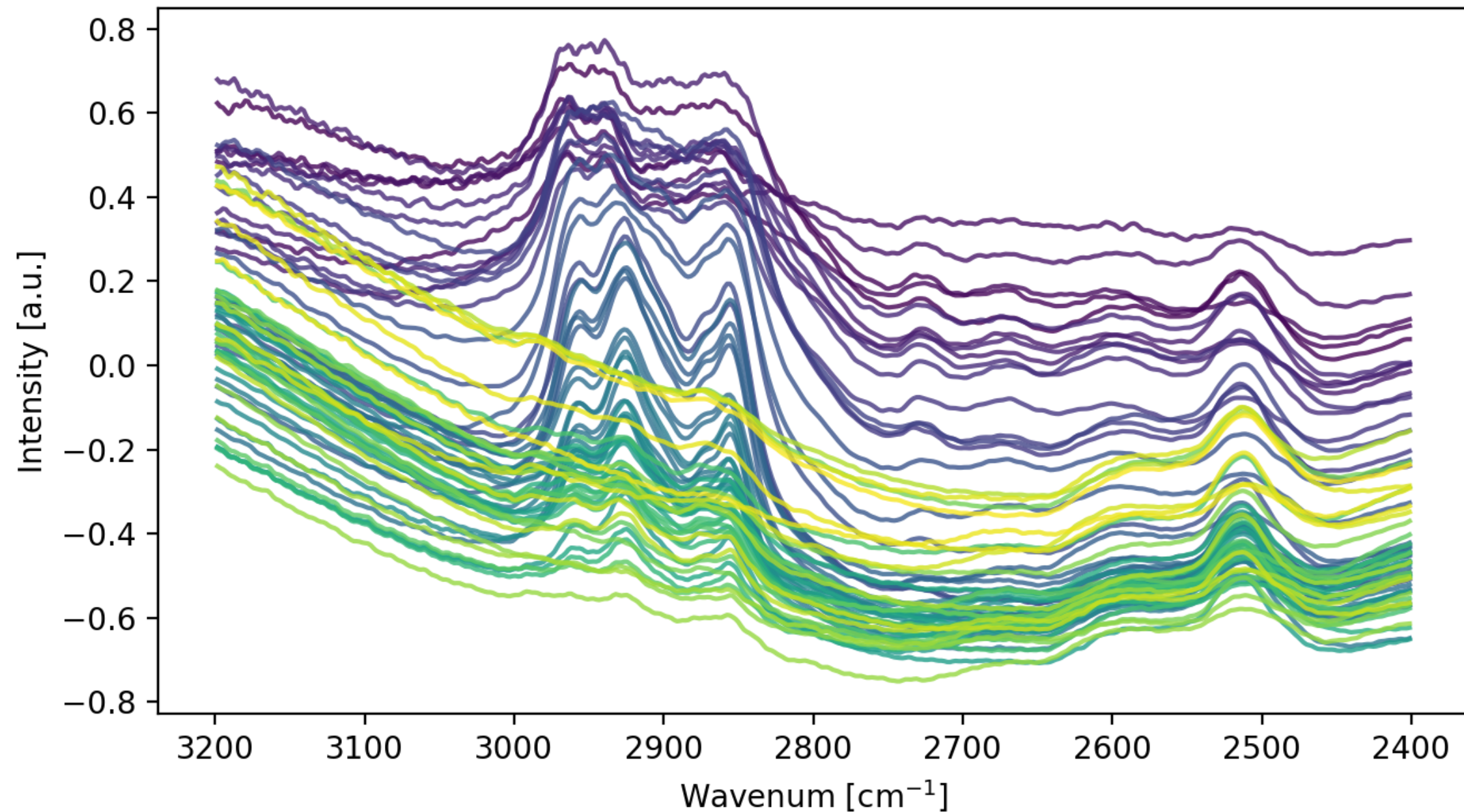
HIGH CARBONATE CONTENT !!!

Results

TEST CASE 4 - Kuwait

CONFUSION MATRIX $\begin{bmatrix} 204 & 0 \\ 25 & 125 \end{bmatrix}$

- Project Details
 - Kuwaiti oil fields
 - Contaminant: Weathered Crude Oil
 - Large scale, multi-year project
- Model: Carbonate
- Number of samples: 354
 - Class A: 204 (>1000)
 - Class B: 150 (<1000)
- Classifier: Gradient Boost
 - Accuracy: 93%
 - F1 Score: 0.91
 - MCC: 0.86

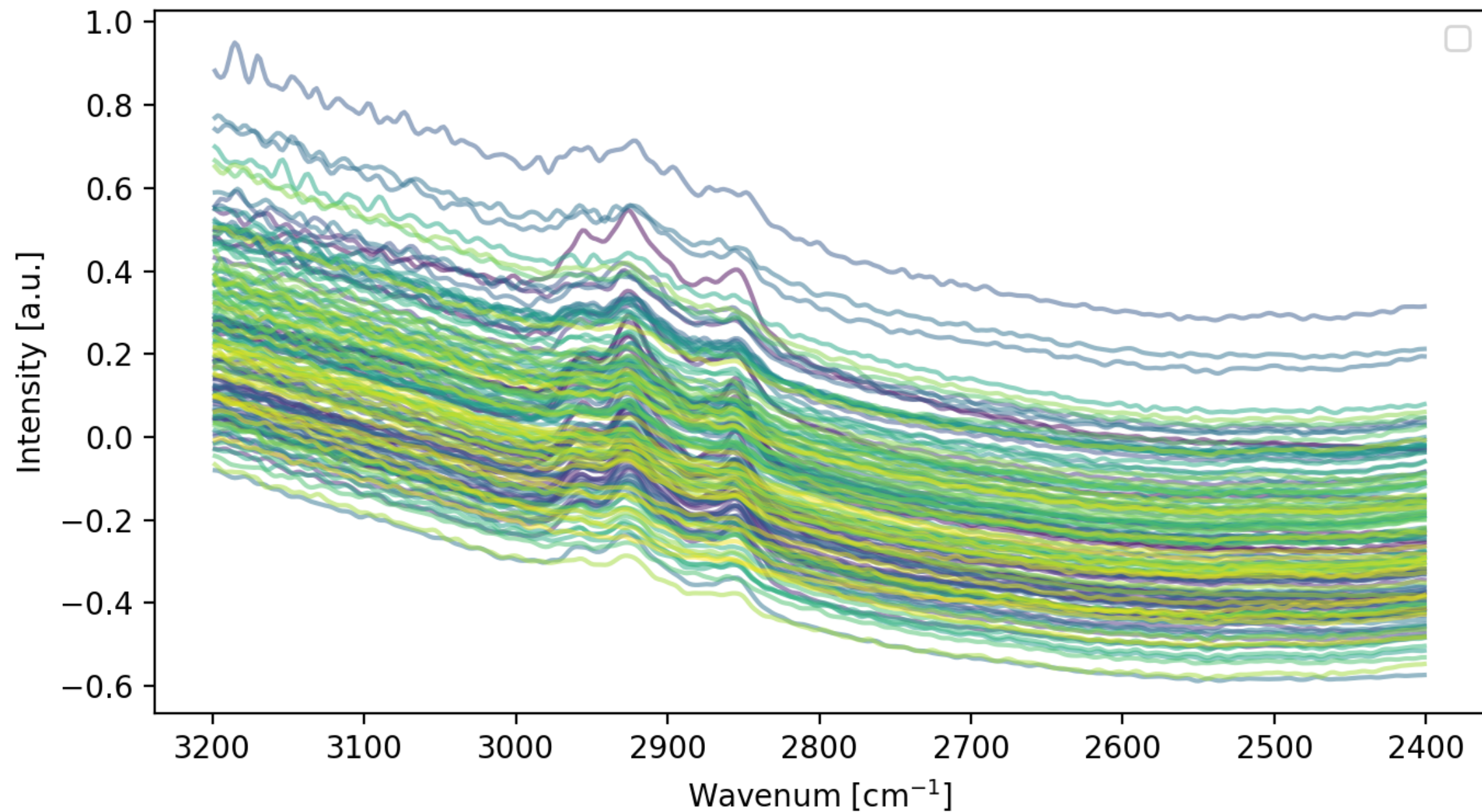


Results

TEST CASE 5 - Antarctica

CONFUSION MATRIX $\begin{bmatrix} 275 & 10 \\ 60 & 392 \end{bmatrix}$

- Project Details:
 - Research facility
 - Contaminant: Antarctic Diesel
 - One week
- Model: Non-carbonate
- Number of samples: 737
 - Class A: 452 (>1000)
 - Class B: 285 (<1000)
- Classifier: SVM
 - Accuracy: 91%
 - F1 Score: 0.92
 - MCC: 0.81

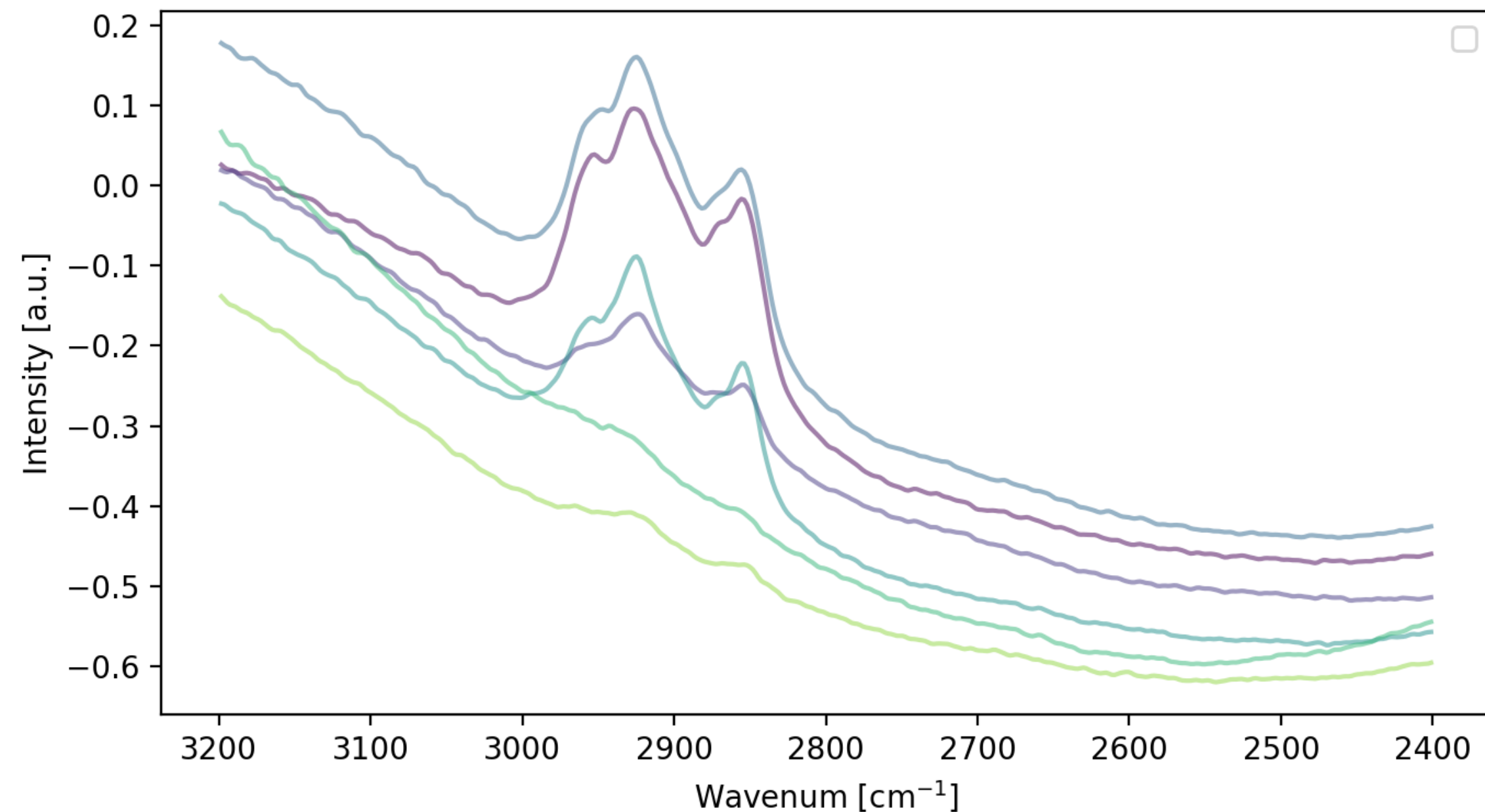


Results

TEST CASE 6 - Metro Adelaide

CONFUSION MATRIX $\begin{bmatrix} 20 & 0 \\ 0 & 12 \end{bmatrix}$

- Project Details:
 - Industrial site
 - Contaminant: Diesel
 - Small scale, several weeks
- Model: Non-carbonate
- Number of samples: 32
 - Class A: 20 (>1000)
 - Class B: 12 (<1000)
- Classifier: SVM
 - Accuracy: 100%
 - F1 Score: 1
 - MCC: 1



USE CASE

PETROL STATION CLEANUP

- Decommissioned petrol stations require cleanup
- Typical backfill soils:
 - Dolomite
 - Sand
- Clean soil samples collected from suppliers.
- Spiked with a known level of contaminant.
- Binary classifiers used to access sample contamination.

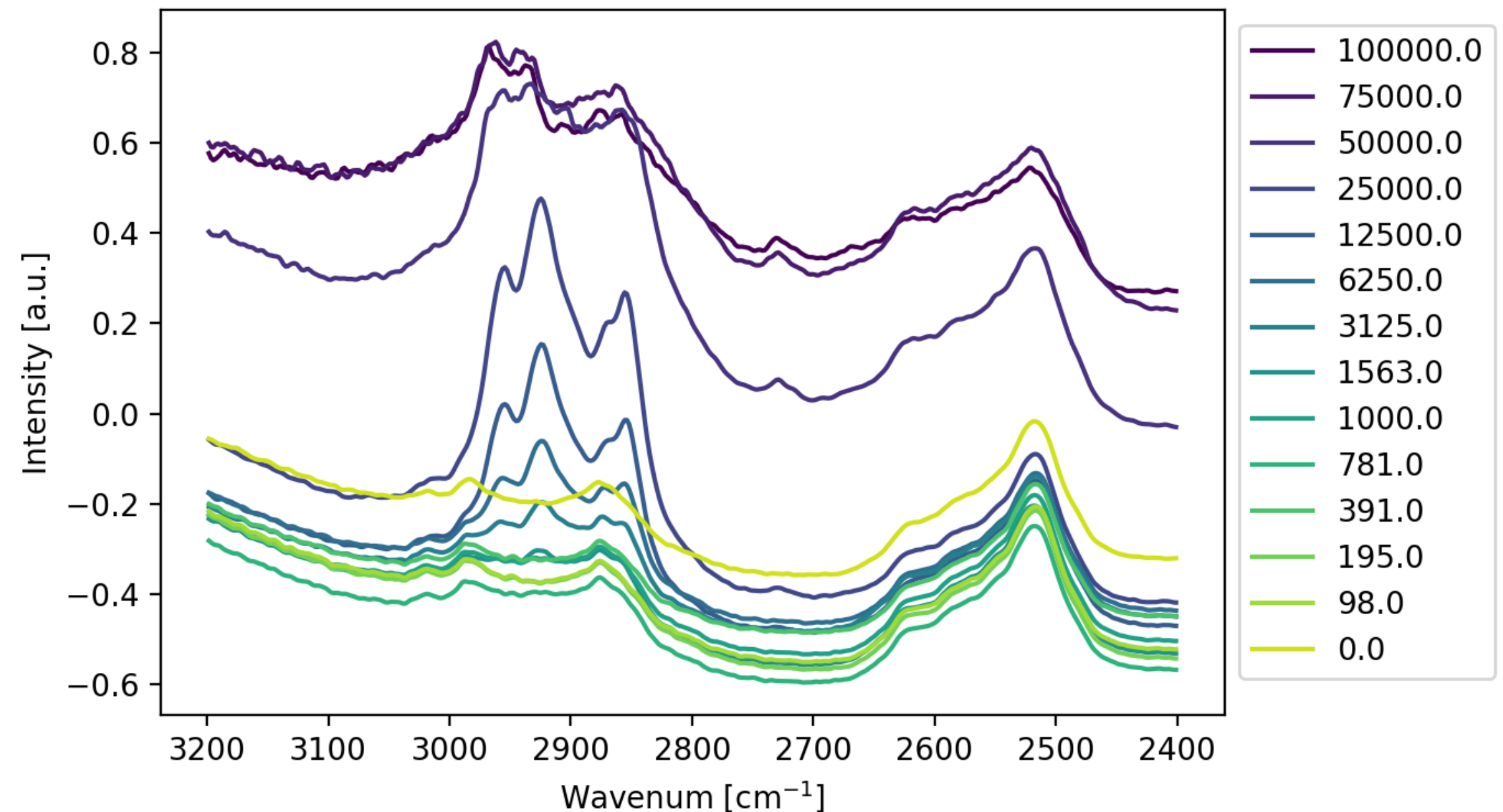


USE CASE

Calibration series: DOLOMITE

CONFUSION MATRIX $\begin{bmatrix} 75 & 15 \\ 0 & 90 \end{bmatrix}$

- Sample details
 - Contaminant: Diesel
 - Calibration standard
- Model: Carbonate
- Number of samples: 180
 - Class A: 90 (>1000)
 - Class B: 90 (<1000)
- Classifier: Gradient Boost
 - Accuracy: 92%
 - F1 Score: 0.93
 - MCC: 0.85

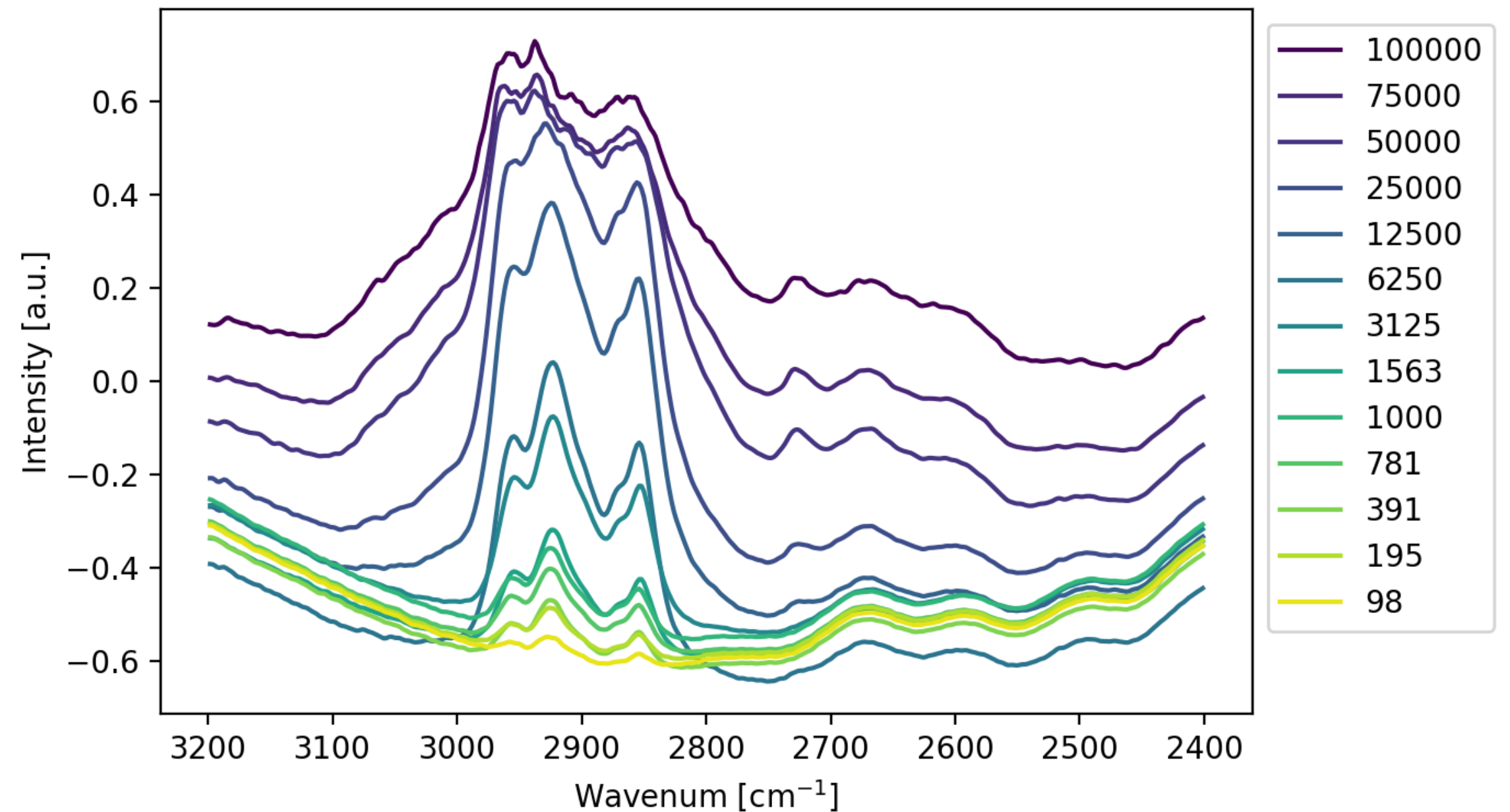


USE CASE

Calibration series: SAND

CONFUSION MATRIX $\begin{bmatrix} 90 & 0 \\ 0 & 60 \end{bmatrix}$

- Sample details
 - Contaminant: Diesel
 - Calibration standard
- Model: Non-Carbonate
- Number of samples: 150
 - Class A: 90 (>1000)
 - Class B: 60 (<1000)
- Classifier: SVM
 - Accuracy: 100%
 - F1 Score: 1.0
 - MCC: 1.0



Conclusion

- Developed binary classifier for rapid-assessment of hydrocarbon-contaminated soils.
- Potential issues identified
 - Calcium Carbonate
 - Soil Organic Carbon
 - Misclassified samples around threshold
- Developed a robust method for handling carbonate signatures.
- Organic carbon signature unlikely to be an issue for soils with $\text{SOC} \leq 10\%$.
- Classifier performance assessed on historical customer data with Diesel Range Organics and heavier contaminants.
- Results are promising with prediction accuracy around 90%.
- Future work will involve
 - Training data refinement
 - Refining model structure to catch edge cases
 - Testing lighter hydrocarbons (e.g., Gasoline)

ZILTEK TEAM



**Dr. Sean
Manning**
CEO



Dr. Georgios Tsiminis
CTO



Aubrey Smith
HR



Chaya Smith
IR Scientist



Zhiran Zheng
Multi-media Specialist



**Dr. Krish
Nachimuthu**
Software Engineer



**Deeksha
Beniwal**
Data Scientist



Thank You

QUESTIONS?