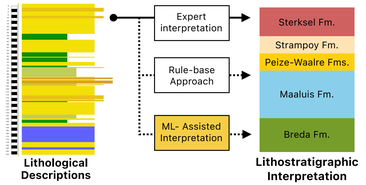
Assessing the Geological Plausibility of Machine Learning Borehole Interpretations

Sebastián Garzón[1], Willem Dabekaussen[2], Eva De Boever[2], Freek Busschers[2], Siamak Mehrkanoon[3], Derek Karssenberg[1]

j.s.garzonalvarado@uu.nl

1 Department of Physical Geography, Faculty of Geosciences, Utrecht University, Utrecht, The Netherlands
2 TNO - Geological Survey of the Netherlands, Utrecht, The Netherlands
3 Department of Information and Computing Sciences, Faculty of Science, Utrecht University, Utrecht, The Netherlands

Introduction



Machine learning (ML) offers an alternative to expert and rule-based **interpretation** of **borehole data** into **stratigraphic units**. While ML predictions may perform well on traditional metrics (e.g. accuracy), they often require post-processing to adhere to **geological principles**. In this work, we propose and test metrics that capture **geological plausibility** to support **model selection**.

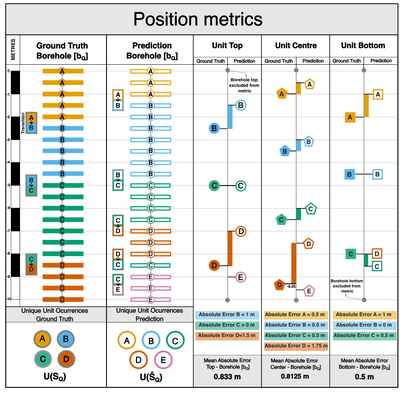
Objectives

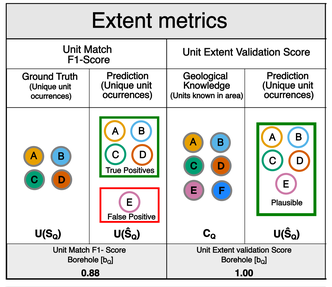
1. Propose a set of **evaluation metrics** to assess the **geological plausibility** of ML predictions.
2. Test the metrics in a case study to evaluate their performance **compared** to **traditional metrics**.

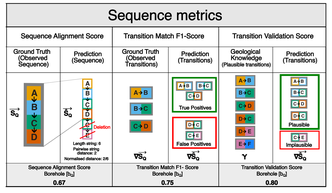
Methodology

1. METRICS DEVELOPMENT

We introduce three sets of **domain-specific evaluation metrics** (**'geology-informed metrics'**) focused on the geological plausibility of **lithostratigraphic unit predictions**. Predictions are assessed based on **unit position**, **geographical extent**, and **sequence**.

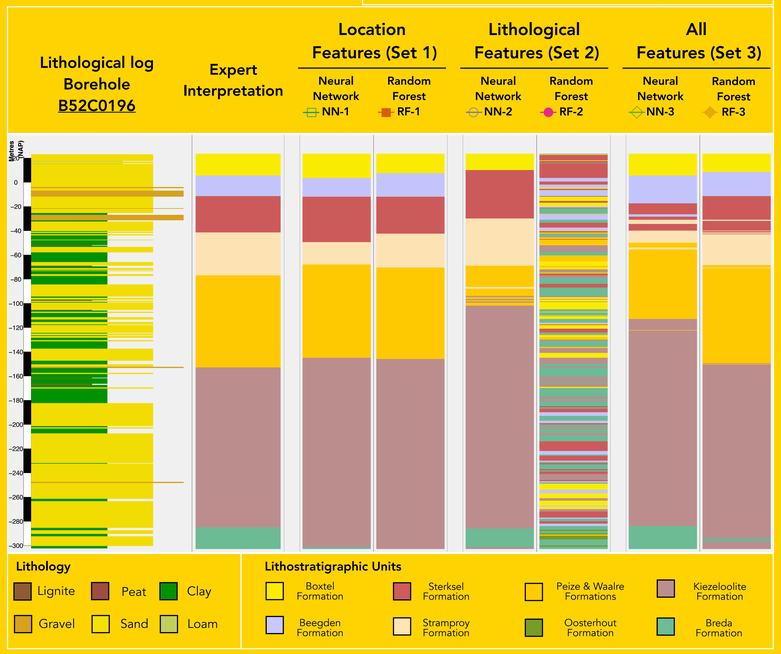






High model performance does not guarantee **geological plausibility.** Domain-specific metrics are essential for **validating** subsurface predictions.

Single borehole interpretation



Interpretation evaluation (Borehole: B52C0196)

	Traditional metrics				Geology-informed metrics							
					Extent metrics		Sequence metrics			Position metrics		
	Accuracy (Acc)	Cohen's Kappa	F1-Score (F1)	Weighted F1 - Score (W-F1)	Unit Match F1-Score (UM-F1)	Unit Extent Validation Score (UEVS)	Sequence Alignment Score (SAS)	Transition Match F1-Score (TM-F1)	Transition Validation Score (TVS)	MAE Top (m)	MAE Centre (m)	MAE Bottom (m)
NN-1	**0.86**	**0.81**	**0.76**	**0.84**	1	1	**1**	**1**	**1**	7.3	4	7.3
RF-1	**0.89**	**0.86**	**0.79**	**0.87**	0.92	1	**0.86**	**0.91**	**1**	3.4	2.8	3.4
NN-2	0.70	0.58	0.61	0.65	0.92	1	0.33	0.57	0.64	20.2	14.3	19
RF-2	0.17	0.05	0.14	0.17	0.85	0.85	0.04	0.21	0.33	48.7	78.9	188.1
NN-3	0.76	0.66	0.76	0.73	1	1	0.46	0.75	0.71	14.1	10.4	11.9
RF-3	**0.91**	**0.87**	**0.82**	**0.89**	1	1	**0.35**	**0.75**	**0.63**	5.7	3	5
	Metric Score [0 = No match] [1 = Perfect match]									Metric Score [0 = Perfect match]		

Utrecht University

TNO Geologische Dienst Nederland

uu.nl/

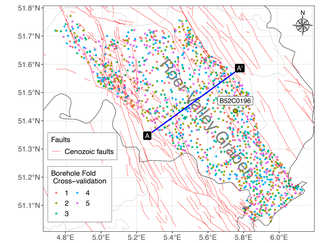
2. CASE STUDY:

We trained a **Random Forest (RF)** and a **Neural Network (NN)** to predict lithostratigraphic units using **1.400 standardised lithological descriptions** of boreholes in the **Roer Valley Graben**. We used three different sets of features as predictors.

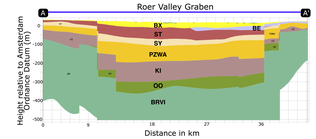
SET	Set Name	# Features	Example
1	Location features	3	Latitude, Longitude & Depth
2	Lithological features	25	Main lithology, colour, grain size
3	All features	28	Location + Lithological Feature Sets

We performed **5-fold cross-validation** and evaluated different hyperparameters for six model configurations: two models (RF and NN) with three feature sets (1, 2, and 3). The **hyperparameters** tested included 'mtry' for **RF** and learning rate, number of LSTM units, and attention layer heads for **NN**.

Study Area



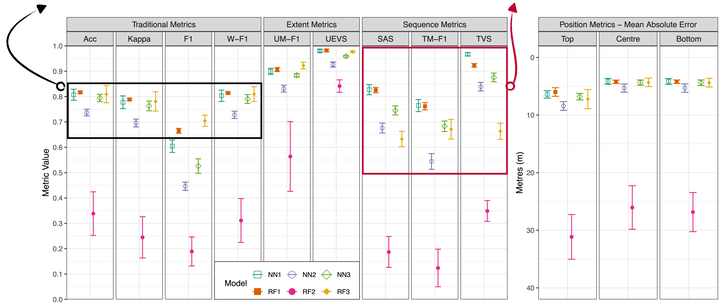
Geologic cross-section



Results (5-fold cross-validation)

- Slight differences
- RF-3 'best model'

- Significant differences
- RF-3 is no longer the 'best model'



Comparison of metrics across 5-fold cross-validation. Each point shows the mean value of the **best hyperparameter configuration** per Set (1, 2, or 3) and method (Random Forest: RF, Neural Network: NN). Error bars represent one standard deviation above and below the mean.

Conclusions

1. **Geology-informed metrics** reveal model **performance differences** that traditional metrics miss.
2. Using **domain-specific evaluation** improves **model selection** for subsurface modelling.

Future steps

Incorporate metrics during training

Integrate the proposed evaluation metrics directly into the **model's loss function** to guide learning toward more **geologically plausible** outcomes that adhere to known constraints (e.g. stratigraphic sequence).

Broaden spatial coverage

TNO– Geological Survey of the Netherlands maintains ~600.000 boreholes, with expert interpretations for ~5%. **ML** can **enhance** or **replace** the existing (time-consuming) rule-based methods.

References

† Selected figures adapted from **Garzón et al. (under review, 2025)** - Assessment of automated stratigraphic interpretations of boreholes with geology-informed metrics.