Computers & Geosciences 207 (2026) 106043

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

Research paper

Assessment of automated stratigraphic interpretations of boreholes with geology-informed metrics

Sebastián Garzón [a,b,*], Willem Dabekaussen [b], Freek S. Busschers [b], Eva De Boever [b], Siamak Mehrkanoon [c], Derek Karssenberg [a]

[a] Department of Physical Geography, Faculty of Geosciences, Utrecht University, Princetonlaan 8A, 3584 CB, Utrecht, The Netherlands
[b] TNO – Geological Survey of the Netherlands, Princetonlaan 6, 3584 CB, Utrecht, The Netherlands
[c] Department of Information and Computing Sciences, Faculty of Science, Utrecht University, Princetonplein 5, 3584 CC, Utrecht, The Netherlands

ARTICLE INFO

Dataset link: https://www.doi.org/10.5281/zenodo.14859951

Keywords:
Automated stratigraphic interpretation
Evaluation metrics
Borehole descriptions
Subsurface modelling

ABSTRACT

Stratigraphic interpretation of borehole data is a fundamental aspect of subsurface geological models, providing critical insights into the distribution of stratigraphic units. However, expert interpretation of all available borehole data is impractical for large-scale regional mapping involving thousands of boreholes. Automated interpretations using machine learning models can significantly increase the number of boreholes included in subsurface geological models. Nevertheless, these predictions must adhere to strict spatial and stratigraphic relationships (e.g. superposition) to ensure geological plausibility, which often requires post-processing tasks. Traditional evaluation metrics commonly used for general-domain classification tasks (e.g. accuracy, F1-score) do not necessarily reflect the geological plausibility of predictions, as they fail to account for the sequential nature and spatial relationships inherent in borehole interpretation. To address this limitation, we propose and evaluate a set of geology-informed metrics that focus on three key aspects of stratigraphic interpretation, namely the expected geographical extent of units (extent metrics), their sequential relationships (sequence metrics), and their vertical positioning along boreholes (position metrics). Using a dataset of 1394 boreholes from the Cenozoic Roer Valley Graben (southeast Netherlands), which covers ~3000 km^2 and includes 15 lithostratigraphic units, we demonstrate that Random Forest and Neural Network models with similar performance on traditional metrics (e.g. accuracy, Cohen's kappa, and F1-score) can differ significantly in their ability to produce geologically plausible predictions. For example, while many model configurations achieve ~75%–80% agreement between expected and predicted classes, the Neural Network models better capture the sequential stratigraphic relationships expected in the study area. Our results underscore the need for domain-specific metrics that offer a more accurate and interpretable assessment of model performance.

1. Introduction

Subsurface modelling requires collaboration between geoscientists and modelling specialists to ensure that large geological datasets are integrated in accordance with fundamental geological principles (Stumpf et al., 2021). In subsurface geological modelling, labelling of borehole intervals is a primary step for constructing layer models and is crucial for subsequent modelling stages (Thomason and Keefer, 2021; Stafleu et al., 2025). However, expert manual labelling of borehole intervals within a common stratigraphic framework is impractical and time-consuming for regional mapping efforts involving thousands of boreholes. Many automated approaches have been developed to support borehole interpretation to predict geologically meaningful categories along boreholes, such as lithostratigraphic units and lithofacies, whose vertical order respects stratigraphic constraints. These methods incorporate stratigraphic relationships and geological knowledge either during prediction (e.g. Markov chains) (Yin et al., 2022; Eidsvik et al., 2004), through rule-based enforcement (Stafleu et al., 2025), or via post-processing (Fullagar et al., 2004; Tokpanov et al., 2020; Wedge et al., 2019). While such methods can increase data density for 3D subsurface models, as in the Dutch GeoTOP model (Stafleu et al., 2011), which integrates 20 times more boreholes than the Digital Geological Model (Gunnink et al., 2013), they can result in complex and time-consuming workflows and post-processing tasks, which can account for up to 50% of the model construction time (Stafleu et al.,

* Corresponding author at: Department of Physical Geography, Faculty of Geosciences, Utrecht University, Princetonlaan 8A, 3584 CB, Utrecht, The Netherlands.
E-mail addresses: j.s.garzonalvarado@uu.nl (S. Garzón), willem.dabekaussen@tno.nl (W. Dabekaussen), freek.busschers@tno.nl (F.S. Busschers), eva.deboever@tno.nl (E. De Boever), s.mehrkanoon@uu.nl (S. Mehrkanoon), d.karssenberg@uu.nl (D. Karssenberg).

https://doi.org/10.1016/j.cageo.2025.106043
Received 17 February 2025; Received in revised form 25 August 2025; Accepted 27 August 2025
Available online 10 September 2025

2021). Machine learning (ML) models are increasingly adopted to support and streamline borehole interpretation in subsurface modelling for both regression and classification tasks (Bhattacharya, 2021). These models provide a flexible approach to learning from training data, with the potential to scale up automatic geological interpretation across large datasets. However, conducting a geology-oriented evaluation of ML-generated predictions, informed by expert knowledge, remains challenging.

The evaluation of machine learning model outputs often relies on general-purpose ML metrics that fail to capture critical aspects of geological problems. For instance, classification models utilise metrics derived from the confusion matrix (Grandini et al., 2020) and regression tasks employ mean absolute error (MAE) or root mean squared error (RMSE) (Botchkarev, 2019). In both cases, borehole interpretations are evaluated pointwise (i.e. at each depth) ignoring the sequential and hierarchical relationships that are fundamental to geological problems. Studies automating the labelling of geological units across dozens to hundreds of boreholes, such as those by Qi and Carr (2006), Tokpanov et al. (2020), and Yang et al. (2023), have achieved ~80%–90% accuracy in the predicted labels. However, even when model predictions match the ground truth, post-processing is necessary to ensure predictions respect basic stratigraphic relationships. These examples underscore the need for evaluation methods that extend beyond label agreement to assess alignment with geological principles and expert knowledge.

Metrics for general-purpose classification tasks, referred to here as traditional metrics, are widely used to evaluate the performance of ML models in automated borehole labelling. However, some studies incorporate domain-specific metrics, highlighting the need to evaluate geological aspects of the prediction. Domain-specific metrics include the Edit Distance (Zhou et al., 2019) and the MAE of Formation Tops (Tokpanov et al., 2020), designed to assess the sequential aspect and positional accuracy of stratigraphic boundaries. In general, these metrics are used during model validation on unseen data to assess performance and can also inform decisions about post-processing. For instance, Tokpanov et al. (2020) demonstrate that applying a post-processing step to the predictions of a Convolutional Neural Network reduces the MAE of the predicted top of a formation on the validation set. Despite the improvement in model evaluation, the proposed metrics failed to incorporate some essential aspects for assessing geological plausibility. Moreover, since traditional metrics such as accuracy and F1-score are commonly used during model selection (e.g. for hyperparameter tuning), integrating metrics that reflect geological plausibility could help guide ML models towards geologically consistent solutions.

This study aims to overcome the limitations of traditional classification metrics in evaluating automated borehole labelling by focusing on two research questions. First, what metrics can be developed to quantify the geological plausibility of stratigraphic unit predictions obtained from borehole data? To address this question, we propose novel geology-informed metrics incorporating geological principles, spatial relationships, and stratigraphic order. Second, can these metrics effectively differentiate between machine learning models based on their ability to generate geologically plausible predictions? For this question, we apply the proposed metrics to predictions generated by various standard ML models in the Roer Valley Graben (south-east Netherlands) using 1394 labelled boreholes, assessing their ability to produce plausible predictions. Rather than identifying the best model or workflow for lithostratigraphic labelling, this study aims to suggest and test alternative methods for integrating geological plausibility into model evaluation using geology-informed metrics. Through this approach, we aim to establish a geology-informed framework for evaluating ML-generated interpretations in borehole labelling tasks.

2. Methodology

This section outlines the approach to evaluating model performance in labelling lithostratigraphic units (i.e. units defined based on lithological characteristics) from borehole data in the south-east Netherlands. First, we describe the case study, comparing two model architectures, Random Forest and Neural Network, using various hyperparameter configurations, three feature sets, and a five-fold cross-validation process. We then introduce both traditional and geology-informed metrics to assess model performance.

2.1. Case study: Roer Valley Graben

Study area

The Cenozoic Roer Valley Graben (RVG) is a major NW-SE trending fault-bounded graben system in the southeastern Netherlands (Fig. 1.A). Tectonically related to Mesozoic and Palaeozoic structures, differential subsidence in the RVG began in the Late Oligocene, resulting in a ~1750 m thick stratigraphic sequence (Fig. 1.B) (Geluk et al., 1995). Mapping efforts rely on interpreting borehole data and labelling lithostratigraphic units along boreholes. In the RVG, the Miocene age Breda Formation, the oldest mapped unit in the Digital Geological Model (Gunnink et al., 2013) for this area, reaches depths of up to 1200 m and marks the model's base.

Dataset and preprocessing

This study uses borehole descriptions from the Roer Valley Graben, including lithological descriptions and expert lithostratigraphic interpretations used in the BRO DGM v2.2, (TNO–GDN, 2014; Gunnink et al., 2013) BRO GeoTOP v1.6, (TNO-GDN, 2023; Stafleu et al., 2021), H3O-Roerdalslenk (TNO-GDN et al., 2014) and H3O-De Kempen (TNO-GDN et al., 2017). We selected 1394 boreholes (Fig. 1.A), ~4.6% of the area's total, containing lithological descriptions and expert lithostratigraphic interpretations (i.e. labels). No additional quality thresholds (e.g. minimum depth or completeness) were applied. Borehole depths range from 0.5 m to 0.9 km, with ~50% shallower than 42.5 m. Interval descriptions, initially at irregular depths, were discretised to 0.5-m intervals. The dataset contains 25 features, comprising 22 macroscopic lithological descriptions based on the Standard Drill Description Method (Bosch, 2000) and three location features (Table 1). During preprocessing, numerical features were normalised to have zero mean and unit variance, and categorical features were encoded using one-hot encoding. Missing values in numeric features were imputed using the median, and categorical features using the most frequent value (mode). To ensure rigorous evaluation and avoid information leakage, these transformations were applied within the training data partitions during cross-validation. The classification task involves 15 lithostratigraphic units as target variables (Fig. 1.C), which follow the Netherlands' stratigraphic nomenclature (TNO-GDN, 2020; Hummelman et al., 2019). The units occupy distinct depth ranges, which correspond to their stratigraphic position in the RVG (Fig. 1. D & E). The dataset is highly imbalanced, with the three most abundant classes accounting for ~50% of borehole intervals.

Modelling approach

To illustrate the differences between prediction models across metrics, we selected the Random Forest (RF) method and a Neural Network (NN) architecture. These models were chosen because RF and NN are established techniques used for classification tasks (Fernández-Delgado et al., 2014), and they handle sequential information in distinct ways. For the RF method (Breiman, 2001), we used the implementation from the R package 'ranger' (Wright and Ziegler, 2017). The NN was implemented using the R implementation of TensorFlow (Abadi et al., 2015) and consists of a ragged input layer for variable-length sequences (e.g. variable borehole depth), followed by a bidirectional

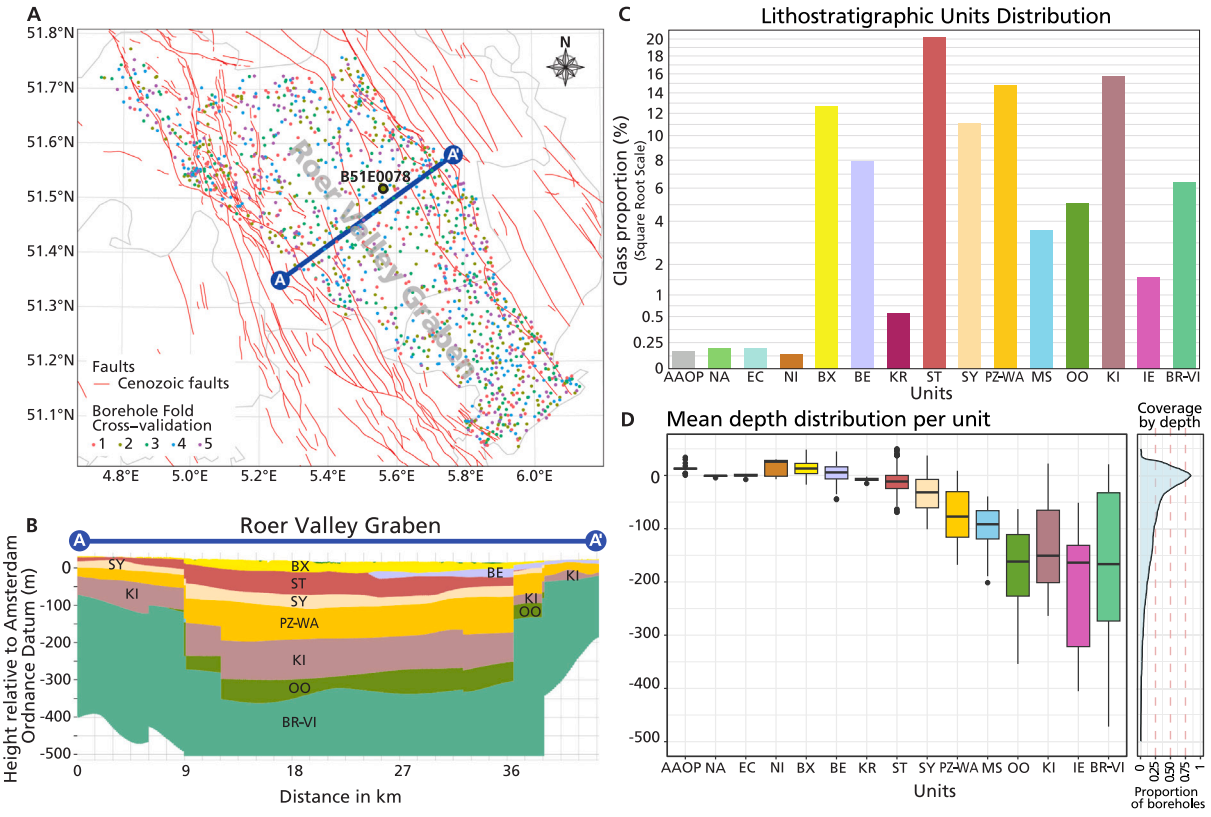

Fig. 1. (A) Map of the Roer Valley Graben area, delimited by Cenozoic faults (v. Gessel et al., 2021), with boreholes grouped by cross-validation fold. (B) SW-NE Cross-section of the Roer Valley Graben, as modelled in the Digital Geological Model (DGM) up to −500 m with respect to the Amsterdam Ordnance Datum (NAP as denoted in Dutch) showing the distribution of lithostratigraphic units in the area. Location of the cross-section indicated in panel A. (C) Proportion of lithostratigraphic units based on boreholes. In the figure, all formal 'NU' (Upper North Sea Group) formation prefixes have been omitted for clarity. AAOP: Made Ground, NA: Naaldwijk Formation, EC: Echteld Formation, NI: Nieuwkoop Formation, BX: Boxtel Formation, BE: Beegden Formation, KR: Kreftenheye Formation, ST: Sterksel Formation, SY: Stramproy Formation, PZ-WA: Peize & Waalre Formations, MS: Maasluis Formation, OO: Oosterhout Formation, KI: Kiezeloölite Formation, IE: Inden Formation, BR-VI: Breda & Ville Formations. TNO-GDN (2020) provides a detailed description of the Formations. (D) Mean depth distribution per lithostratigraphic unit with respect to NAP. (E) Depth-wise proportion of borehole coverage.

Table 1
Predictor variable descriptions and feature sets. SBB-5.1 correspond to the Standard Drill Description Method used by the TNO - Geological Survey of the Netherlands (TNO-GDN) (Bosch, 2000). NAP: Amsterdam Ordnance Datum.

Variable	Description	Feature set		
		Location (1)	Lithology (2)	All (3)
X coordinate	Longitude (EPSG:28992)	X		X
Y coordinate	Latitude (EPSG:28992)	X		X
Depth	Depth with respect to NAP (m)	X		X
Main Lithology	Soil Type [SBB 5.1 - L3.1]		X	X
Main colour	Soil Type [SBB 5.1 - L4.3]		X	X
Presence of plants	Plant fragments [True/False]		X	X
Presence of shells	Shell fragments [True/False]		X	X
Presence of clay chunks	Clay chunks [True/False]		X	X
Sand median	Sand median [SBB 5.1 - L7.2.1]		X	X
Sand median class	Sand median class [SBB 5.1 - L7.2.2]		X	X
Sand median class*	Sand median class (Preprocessed)		X	X
Lime content	Lime Content [SBB 5.1-L4]		X	X
Clay mixture	Clay admixture [SBB 5.1-L3.3.1]		X	X
Silt admixture	Silt admixture [SBB 5.1-L3.3.2]		X	X
Sand admixture	Sand admixture [SBB 5.1-L3.3.3]		X	X
Gravel admixture	Gravel admixture [SBB 5.1-L3.3.4]		X	X
Humus admixture	Humus admixture [SBB 5.1-L3.3.5]		X	X
Shell material	Shell Material [SBB 5.1-L12.2]		X	X
Consistency	Consistency [SBB 5.1]		X	X
Plant residue	Plant residue [SBB 5.1 - L11.1]		X	X
Organic material	Organic material		X	X
Shell material	Shell material		X	X
Percentage of clay	Clay percentage		X	X
Mica residue	Amount of mica [SBB 5.1 - L13.1]		X	X
Glauconite residue	Amount of Glauconite [SBB 5.1 - L13.2.2]		X	X

Table 2
Configuration of hyperparameters for the Random Forest and Neural Network models.

Model configuration			Hyperparameters				Number of setups
Model type	Feature set	Model ID	Random Forest	Neural Network			
			Features sampled per split (mtry)	Learning rate	LSTM Units	Attention layer (Heads)	
Random Forest	(1) Location	RF1	1, 2, 3	N/A	N/A	N/A	3
	(2) Lithology	RF2	10, 20, 30, 40, 50, 60, 70, 80	N/A	N/A	N/A	8
	(3) All	RF3	10, 20, 30, 40, 50, 60, 70, 80	N/A	N/A	N/A	8
Neural Network	(1) Location	NN1	N/A	0.01, 0.001	16, 32, 64	1, 4, 8	18
	(2) Lithology	NN2	N/A	0.01, 0.001	16, 32, 64	1, 4, 8	18
	(3) All	NN3	N/A	0.01, 0.001	16, 32, 64	1, 4, 8	18

LSTM layer (Hochreiter and Schmidhuber, 1997), a multi-head attention layer (Vaswani et al., 2017), and a dense output layer with softmax activation. We use categorical cross-entropy as the loss function. This NN architecture is expected to integrate the sequential aspect of the task as it includes elements of recurrent neural networks and attention mechanisms.

We evaluated each model using three input feature sets (Table 1). Set 1 comprised location features, set 2 contained lithological features, and set 3 combined both. The three feature sets were designed to assess the predictive value of spatial location and lithological characteristics, which represent fundamentally different geological information. While location features capture spatial trends and regional stratigraphic variations, lithological features describe local sediment properties. Combining both sets tests whether integrating spatial and compositional data improves predictions. We did not perform formal feature selection, as our goal was to evaluate and compare model performance on complete, geologically meaningful feature groups rather than optimising input variables via statistical methods.

Additionally, we assessed the impact of varying hyperparameters (see Table 2). For RF, we adjusted the number of variables considered at each split when building the decision trees (i.e. the 'mtry' hyperparameter). For the NN, we modified the learning rate, number of LSTM units, and number of heads in the attention layer.

Cross-validation and evaluation

To assess model generalisation, we implemented a five-fold cross-validation. The dataset was split using the stratified group k-fold function from scikit-learn (Pedregosa et al., 2011), which balances the target variables and prevents intervals from the same borehole from appearing in multiple folds, ensuring that no borehole appears in more than one fold. For each iteration, three folds were used for training, one for validation (for NN models), and one for testing. Neural networks were trained for up to 300 epochs, with early stopping implemented after 30 epochs without improvement in validation loss. We trained the Neural Network using the Adam optimiser with batches that have entire boreholes, each containing variable-length sequences handled via the ragged input layer. The batch size was set to 32. Finally, we evaluate the resulting predictions of the NN and RF methods for each set and hyperparameter configuration using both traditional and geology-informed metrics. The RF and NN models generate class scores at each depth, using the proportion of tree votes for RF and softmax confidence scores for NN. The predicted class at each depth is the one with the highest score. We then evaluate the sequence of predicted classes per borehole using traditional and geology-informed metrics.

Seventy-three model setups were evaluated (Table 2) based on the different feature sets and hyperparameter configurations. Individual RF models typically train in under five minutes, while NN training times vary by configuration- depending on the number of features, hyperparameter settings, and whether early stopping was applied- and can take up to an hour.

2.2. Metrics

This study categorises metrics for evaluating model performance into two main types: traditional metrics and geology-informed metrics (Table 3). In our work, traditional metrics refer to standard metrics used in classification tasks, such as accuracy or F1-score. While these are useful for checking individual label predictions, they are not suited for borehole predictions, which require interpreting vertically ordered sequences that follow a stratigraphic structure. These metrics do not account for the geological rationale behind a prediction, such as the expected continuity, transitions, and spatial extent of units. Geology-informed metrics refer to the proposed metrics in this paper, which aim to capture essential geological aspects of model predictions. Both traditional and geology-informed metrics can be used for overall model evaluation and intermediate steps such as hyperparameter tuning. However, no single metric can fully capture all facets of prediction quality. Just as accuracy is complemented by metrics like the F1 score in classification tasks, geology-informed metrics should be applied together to provide a more comprehensive and geologically meaningful assessment.

A key challenge in evaluating lithostratigraphic predictions is the variability and interpretative nature of the ground truth labels. These interpretations are based on manual labelling by an expert -typically without a measurement of uncertainty. Therefore, discrepancies between model predictions and expert labels may not indicate errors but rather reflect alternative geologically plausible interpretations. Recognising this limitation, our geology-informed metrics assess the plausibility of predictions from a broader geological perspective, providing complementary insights beyond direct label matching.

Metrics implementation

For clarity, we introduce the notation that will be used throughout the evaluation of the metrics. Specifically, we define $\mathcal{B} = \{b_1, \ldots, b_N\}$ as the set of boreholes, where N is the total number, and i represents a unique index of each borehole. Each borehole b_i contains n_i data instances, represented as the collection:

$$b_i = \{(d_j, \hat{S}_{i,d_j}, S_{i,d_j}) \mid j = 1, \ldots, n_i\}, \tag{1}$$

where d_j is the depth value, and $\hat{S}_{i,d_j}$ and S_{i,d_j} are the predicted and ground truth classes, respectively, at depth d_j. The set of all possible stratigraphic units (i.e. classes) is denoted by $\mathcal{C} = \{c_1, \ldots, c_K\}$, with K being the total number of classes. For simplicity, we denote $\mathbf{D}_i = (d_1, \ldots, d_{n_i})$, $\hat{\mathbf{S}}_i = (\hat{S}_{i,d_1}, \ldots, \hat{S}_{i,d_{n_i}})$, and $\mathbf{S}_i = (S_{i,d_1}, \ldots, S_{i,d_{n_i}})$ as the ordered vectors for depth, predicted units, and ground truth units, respectively, for b_i.

Traditional metrics

We use Accuracy, Cohen's Kappa, Macro F1, and Weighted F1 to evaluate model performance, considering them traditional metrics widely used in classification tasks (Naidu et al., 2023). These metrics are defined based on a confusion matrix (Grandini et al., 2020), denoted as CM. The $K \times K$ confusion matrix compares predicted and actual

Table 3
Traditional and Geology-informed evaluation metrics for automated stratigraphic interpretation of boreholes.

Category	Evaluation target	Focus	Evaluation metric		
			Metric name	Value range	Geologic rational
Traditional metrics	Overall model fit	Classification agreement	Accuracy Cohen's Kappa	[0,1] [−1,1]	N/A N/A
		Class balance	F1-Score Weighted F1-Score	[0,1] [0,1]	N/A N/A
Geology-informed metrics	Position: Vertical Alignment	Unit top	Mean Absolute Error Top	[0, ∞) (m)	Predicted top units should align with the expected depth of the units
		Unit centre	Mean Absolute Error Centre	[0, ∞) (m)	Predicted centre of units should align with the ground truth, reflecting both the expected vertical position and thickness of the units
		Unit Bottom	Mean Absolute Error Bottom	[0, ∞) (m)	Predicted bottom units should align with the expected depth
	Extent: Geographical range of units	Unit presence	Unit Match F1-Score	[0,1]	Predicted units should correspond to those observed in the borehole
			Unit Extent Validation Score	[0,1]	Predicted units should appear only within their expected geographical extent
	Sequence: Stratigraphic Order	Unit's transitions	Transition Match F1-Score	[0,1]	Predicted transitions should match expected geological transitions in borehole
			Transition Validation Score	[0,1]	Predicted transitions should be geologically plausible, based on known stratigraphic relationships
		Complete sequence	Sequence Alignment Score	[0,1]	Predicted sequences should align with ground truth sequence

class labels, with diagonal entries representing correct predictions. Values in the CM are defined as:

$$\mathrm{CM}_{k,l} = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \mathbb{1}\left(\hat{S}_{i,d_j} = c_k \wedge S_{i,d_j} = c_l\right), \tag{2}$$

where $k, l \in \{1, \ldots, K\}$, with K being the total number of target features (classes), and $\mathbb{1}(\cdot)$ denoting an indicator function.

These metrics evaluate different aspects of the model's performance. Accuracy measures the proportion of correct predictions, Cohen's Kappa considers the likelihood of agreement by chance, and the Macro and Weighted F1 scores assess class-wise performance, with the latter accounting for class distribution. Details on their computation can be found in Grandini et al. (2020).

Proposed geology-informed metrics

We assess model performance based on geological plausibility, using three metric sets — position, extent, and sequence (Table 3) — to evaluate the vertical placement, geographical extent, and sequential relationships of stratigraphic units. Unlike traditional metrics, which compare individual data instances, our approach analyses the entire boreholes as groups of data instances. This methodology aligns with geologists' labelling results interpretation and offers evaluation criteria reflecting geological reasoning.

For this reason, we define operations for ordered vectors and sets, introducing notations necessary for evaluating the sequential nature of stratigraphic unit predictions. The function $U(\cdot)$ identifies unique elements in an ordered vector, such that for $\mathbf{S}_i$, $U(\mathbf{S}_i) = \{S_{i,d_j} \mid S_{i,d_j} \neq S_{i,d_k}, \forall j \neq k\}$, where no repetitions occur. The function $\mathrm{Seq}(\cdot)$ denotes the ordered sequence of stratigraphic units in a borehole, expressed as $\mathrm{Seq}(S_i) = \overrightarrow{S_i} = \{S_{i,d_1}\} \cup \{S_{i,d_j} \mid S_{i,d_j} \neq S_{i,d_{j-1}}, j = 2, \ldots, n_i\}$. The sequence $\overrightarrow{S} = (s_1, \ldots, s_g)$ consists of g ordered elements, where non-consecutive repetition is allowed, and the original notation S_{i,d_j} no longer applies to the elements of the sequence. Finally, $\mathrm{Transitions}(\cdot)$ represents the set of transitions between consecutive elements in a sequence, such that $\mathrm{Transitions}(\overrightarrow{S_i}) = \nabla\overrightarrow{S_i} = \{(s_j, s_{j+1}) \mid s_j, s_{j+1} \in \overrightarrow{S}_i, j = 1, \ldots, g-1\}$.

Position metrics

The position metrics for a collection of boreholes $\mathcal{B}$ are calculated by computing the mean absolute error (MAE) between the predicted and ground truth values of the top, centre, and bottom positions of the stratigraphic units in each borehole. For each borehole i, we first define M_i as the intersection of the unique predicted $U(\hat{\mathbf{S}}_i)$ and ground truth $U(\mathbf{S}_i)$ values where comparison is possible. For the MAE-Top metric, M_i excludes the topmost units in both $U(\mathbf{S}_i)$ and $U(\hat{\mathbf{S}}_i)$, while for the MAE-bottom position, M_i excludes the bottommost units. The topmost and bottommost units are excluded because borehole limits are arbitrary and may not align with the top or bottom of a unit. Fig. 2 illustrates this process using an example borehole, comparing the predicted and true sequences alongside the computed top, centre, and bottom position metrics. The metrics are computed as follows:

$$\text{MAE-Top} = \frac{1}{\sum_{i=1}^{N} |M_i|} \sum_{i=1}^{N} \sum_{c \in M_i} |\max(\mathrm{Z}(\hat{S}_{i,c})) - \max(\mathrm{Z}(S_{i,c}))|, \tag{3}$$

$$\text{MAE-Centre} = \frac{1}{\sum_{i=1}^{N} |M_i|} \sum_{i=1}^{N} \sum_{c \in M_i} |\mathrm{median}(\mathrm{Z}(\hat{S}_{i,c})) - \mathrm{median}(\mathrm{Z}(S_{i,c}))|, \tag{4}$$

$$\text{MAE-Top} = \frac{1}{\sum_{i=1}^{N} |M_i|} \sum_{i=1}^{N} \sum_{c \in M_i} |\min(\mathrm{Z}(\hat{S}_{i,c})) - \min(\mathrm{Z}(S_{i,c}))|, \tag{5}$$

with N representing the total number of boreholes in the dataset and c representing a class value in the intersection set M_i. $|M_i|$ is the cardinality of the intersection set. $Z(S_{i,c})$ and $Z(\hat{S}_{i,c})$ represent the set of depth values for the cth unit in borehole b_i. Specifically, $Z(S_{i,c}) = \{d_j \mid S_{i,d_j} = c, j = 1, \ldots, n_i\}$, where n_i is the number of depth values in borehole b_i.

Geographical extent metrics

We implemented two metrics to evaluate whether predictions reflect the spatial extent of geological units defined by deposition processes and geological structures (Fig. 3).

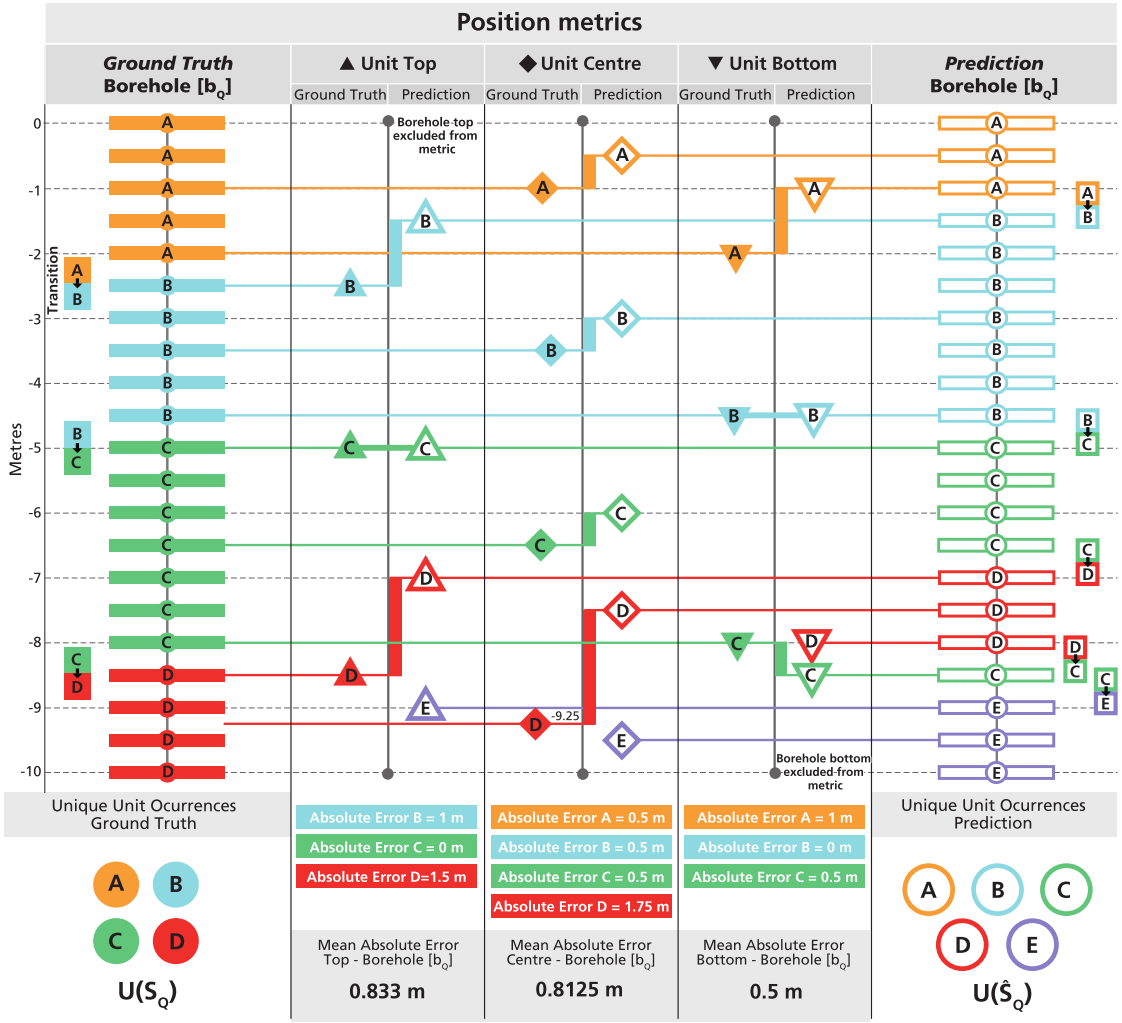

Fig. 2. Position metrics calculation for Borehole b_Q. The Ground Truth S_Q and Prediction $\hat{S}_Q$ shows values of an example borehole with five geological units (A, B, C, D, and E) every 0.5 m. For the ground truth S_q (left), the sequence contains the unique units $U(S_q) = \{A, B, C, D\}$, the ordered sequence $\overrightarrow{S_q} = (A, B, C, D)$ and the transitions $\nabla\overrightarrow{S_q} = \{(A, B), (B, C), (C, D)\}$. For the prediction $\hat{S}_q$ (right), the sequence contains the unique units $U(\hat{S}_q) = \{A, B, C, D, E\}$, the ordered sequence $\overrightarrow{\hat{S}}_q = (A, B, C, D, C, E)$ and the transitions $\nabla\overrightarrow{\hat{S}}_q = \{(A, B), (B, C), (C, D), (D, C), (C, E)\}$. The middle panel illustrates the computed position metrics for each unit in the sequence: unit top, unit centre, and unit bottom, which quantify the positional differences between the predicted and true occurrences of each unit in the borehole. For notation and definitions, see Section 2.2.

The first metric, the Unit Match - F1 Score (UM-F1), evaluates the agreement between the unique predicted units $U(\hat{S})$ and the ground truth units $U(S)$. The UM-F1 score is computed as follows:

$$\text{UM-F1} = \frac{1}{N}\sum_{i=1}^{N}\frac{2|U(\hat{S}_i)\cap U(S_i)|}{2|U(\hat{S}_i)\cap U(S_i)| + |U(\hat{S}_i)\backslash U(S_i)| + |U(S_i)\backslash U(\hat{S}_i)|}. \quad (6)$$

The second metric, the Unit Extent Validation Score (UEVS), assesses the alignment of predicted formations with the expected extent derived from external geological maps. Unlike UM-F1, UEVS does not rely on ground truth data, but instead compares with established geological knowledge of the area. The UEVS is computed as:

$$\text{UEVS} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\text{U}(\hat{S}_i)\cap C_i|}{|\text{U}(\hat{S}_i)|}, \quad (7)$$

with C_i representing the set of expected geological units in borehole B_i based on external sources.

Sequence metrics

The sequence metrics assess the model's ability to predict the correct stacking order of geological units along a borehole, and whether these predictions align with the observed or expected stratigraphic relationships (Fig. 3). For this, we define the sequence of unique units along each borehole from top to bottom.

The first metric, the Sequence Alignment Score (SAS), measures the overall similarity between the predicted and ground truth sequences based on the Optimal String Alignment (OSA) algorithm (Loo, 2014) denoted as OSA(). This algorithm calculates the minimum number of edits needed to align two sequences (i.e. insertions, deletions, substitutions, and adjacent transpositions).

The SAS is calculated as follows:

$$\text{SAS} = \frac{1}{N}\sum_{i=1}^{N} 1 - \frac{\text{OSA}(\overrightarrow{S}_i, \overrightarrow{\hat{S}}_i)}{\max(\{|\overrightarrow{S}_i|, |\overrightarrow{\hat{S}}_i|\})}, \quad (8)$$

where $|\cdot|$ denotes the length (cardinality) of the sequence, and max() the maximum value of a set.

Next, we introduce the Transition Match F1-Score (TM-F1), which evaluates the model's ability to predict correct transitions between units based on the ground truth. The TM-F1 is calculated as follows:

TM-F1

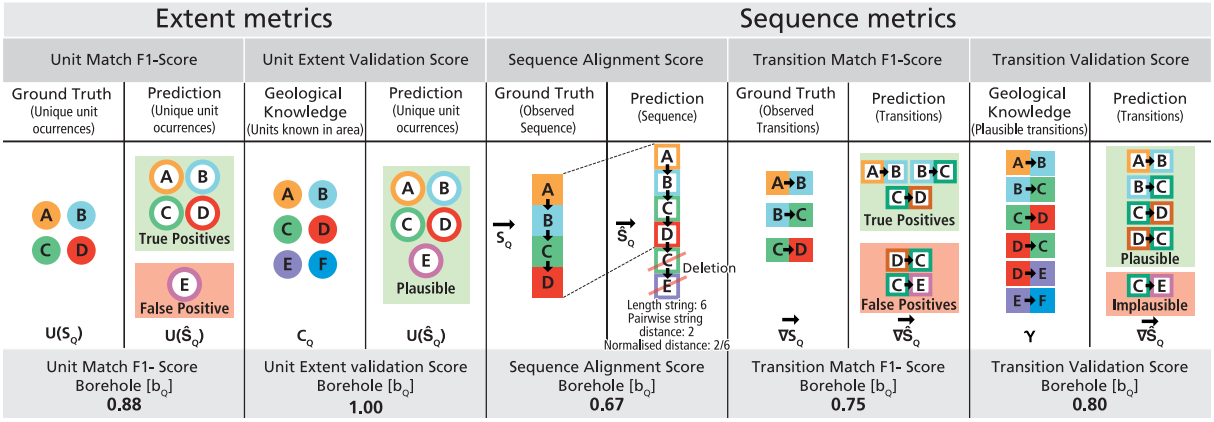

Fig. 3. Extent and sequence metrics calculation for Borehole b_Q. This figure's ground truth and prediction sequences correspond to the example shown in Fig. 2. External geological information for borehole b_Q includes the set of known units in area, $C_Q = \{A, B, C, D, E, F\}$ and the set of plausible transitions between units, $Y = \{(A, B), (B, C), (C, D), (D, C), (D, E), (E, F)\}$. For notation and definitions, see Section 2.2.

Table 4
Model performance results for Borehole B51E0078, taken from a single test fold (unseen data). For each metric, the best value is marked in bold. Results are based on the hyperparameter configuration with the highest mean accuracy across five-fold cross-validation, selected separately for each model type (RF or NN) and feature set. The borehole shown comes from one of the cross-validation test folds. Accuracy: Accuracy, Kappa: Cohen's Kappa, F1: Macro-F1 Score, W-F1: Weighted F1, UM-F1: Unit Match - F1 Score, UEVS: Unit Extent Validation Score, SAS: Sequential Alignment Score, TM-F1: Transition Match - F1 Score, TVS: Transition Validation Score, MAE-Top: Mean Absolute Error - Top, MAE-Centre: Mean Absolute Error Centre, MAE-Bottom: Mean Absolute Error Bottom.

Model	Traditional metrics				Extent metrics		Sequence metrics			Position metrics Mean Absolute Error (m)		
	Accuracy	Kappa	F1	W-F1	UM-F1	UEVS	SAS	TM-F1	TVS	Top	Centre	Bottom
NN1	0.76	0.71	**0.66**	0.77	**0.93**	**1.00**	**0.88**	**0.92**	**1.00**	12.00	8.43	8.43
RF1	0.60	0.53	0.54	0.55	0.86	**1.00**	0.86	0.83	**1.00**	**6.50**	12.25	12.25
NN2	0.60	0.52	0.53	0.51	0.83	**1.00**	0.26	0.50	0.61	**6.50**	15.30	20.11
RF2	0.37	0.23	0.26	0.38	0.86	0.86	0.06	0.07	0.40	80.90	61.21	67.90
NN3	**0.77**	0.72	0.64	0.78	**0.93**	**1.00**	0.50	0.71	0.69	19.17	11.25	11.07
RF3	**0.77**	**0.73**	**0.66**	**0.80**	**0.93**	**1.00**	0.44	0.71	0.73	13.92	**7.79**	**7.84**

$$= \frac{1}{N}\sum_{i=1}^{N} \frac{2|U(\nabla\overrightarrow{\hat{S}_i}) \cap U(\nabla\overrightarrow{S_i})|}{2|U(\nabla\overrightarrow{\hat{S}_i}) \cap U(\nabla\overrightarrow{S_i})| + |U(\nabla\overrightarrow{\hat{S}_i}) \setminus U(\nabla\overrightarrow{S_i})| + |U(\nabla\overrightarrow{S_i}) \setminus U(\nabla\overrightarrow{\hat{S}_i})|}. \tag{9}$$

Finally, the Transition Validation Score (TVS) evaluates predicted transitions against a predefined set of observed transitions (Y) based on geological knowledge of the study area. Transitions outside this set are considered geologically implausible.

$$\text{TVS} = \left(\sum_{i=1}^{N} \left|U(\nabla\overrightarrow{\hat{S}_i}) \cap Y\right|\right) \cdot \left(\frac{1}{\sum_{i=1}^{N} \left|U(\nabla\overrightarrow{\hat{S}_i})\right|}\right) \tag{10}$$

3. Results

In this section, we describe two main aspects of the analysis. First, we examine the overall performance of the Random Forest (RF) and Neural Network (NN) models. We focus on differences between model types (RF, NN) and feature sets (1, 2, and 3, Table 1). Second, we evaluate the impact of different hyperparameter configurations for each model using traditional and geology-informed metrics.

Before comparing model performance, we outline the process for assigning lithostratigraphic units by depth. In RF models, the predicted unit is the class with the highest vote share among trees. In NN models, it is the class with the highest softmax activation. We treat softmax outputs as vote proportions, reflecting the model's relative support for each class. Though not actual probabilities, this normalised score can be viewed as a confidence distribution over classes. Fig. 4 shows this process for two predictions from a single borehole.

To illustrate the performance differences between different models on unseen data, we show a prediction for a single borehole from the test set of one of the five cross-validation folds (Fig. 5) and compare traditional and geology-informed metrics (Table 4). For each model (RF or NN) and feature set (Set 1, 2, or 3), hyperparameter configurations were selected based on the highest mean accuracy across five-fold cross-validation. Despite similar traditional metric values for NN1, NN3, and RF3, sequence metrics favour RF1 and NN1 in this single-borehole evaluation. Models using lithological features (Sets 2 and 3) exhibit more implausible transitions, characterised by lower SAS and TVS values. Common misclassifications involve unit interbedding, which is unexpected at the formation scale in the Roer Valley Graben. While filtering (e.g. removing single-occurrence units) can reduce some errors, others require expert review (e.g. the missing Oosterhout Formation in RF1 prediction).

3.1. Overall model performance

Unlike the single-borehole example above, the following analysis evaluates model performance averaged across all five cross-validation folds. The values presented in the text and figures represent the mean and standard deviation calculated from these folds. Key observations are summarised below, with Fig. 6 showing the best-performing models based on their respective optimal hyperparameter configurations for each model (RF or NN) and feature set per metric.

Traditional metrics for model evaluation show comparable performance between RF and NN models, with RF1 and RF3 demonstrating a slight advantage (Fig. 6). The best models achieve an accuracy of 0.82 ± 0.005 (mean ± standard deviation) (RF1), 0.79 ± 0.01 in Cohen's Kappa (RF1), and 0.70 ± 0.02 for F1-score (RF3), indicating good agreement with the ground truth.

While RF1 and RF3 outperform NN models across traditional metrics, the differences are slight and within the standard deviation (e.g.

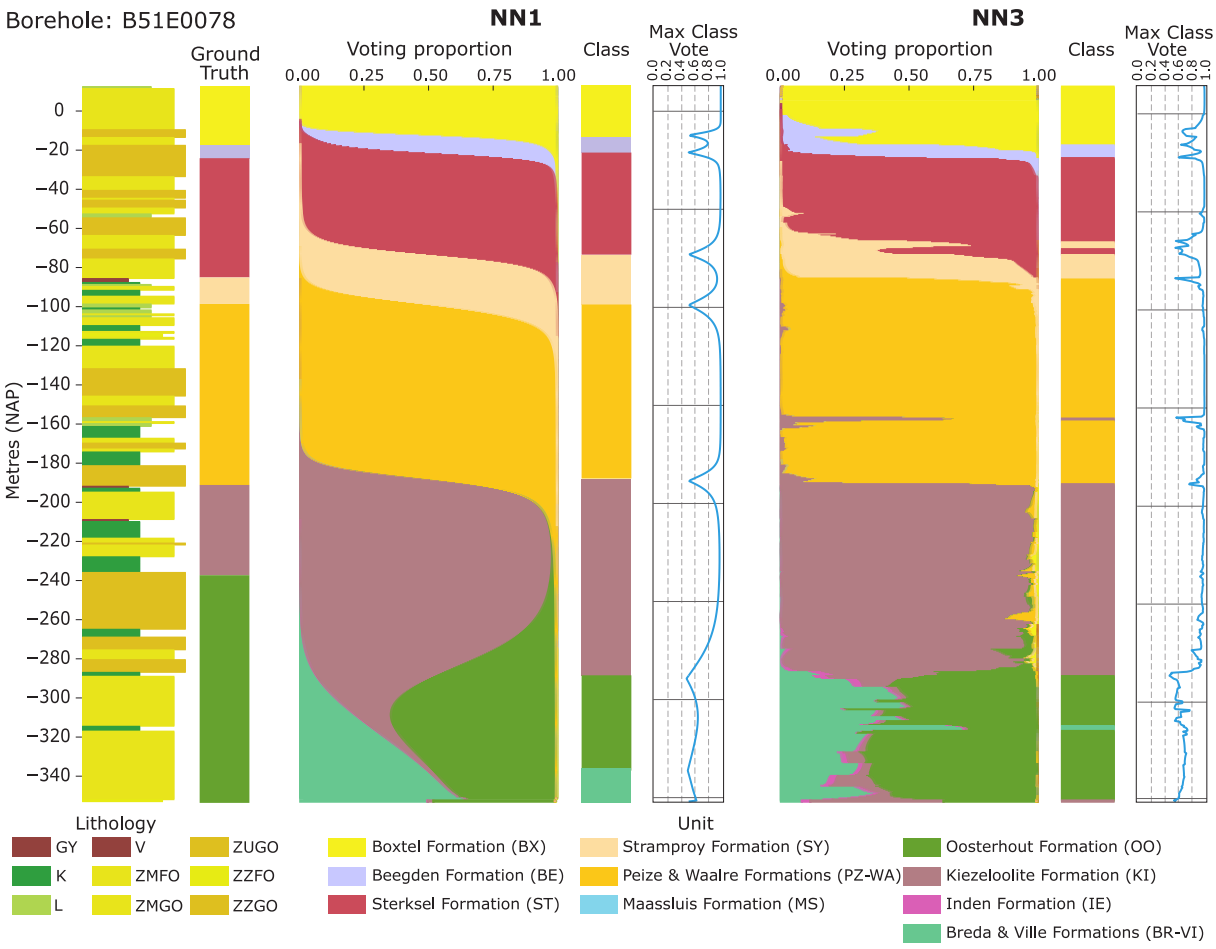

Fig. 4. Voting proportions for lithostratigraphic unit predictions in borehole B51E0078 using Neural Network models NN1 and NN3. NN1 (64 LSTM units, one head in the attention layer, 0.01 learning rate, location features only) and NN3 (64 LSTM units, four heads in the attention layer, 0.001 learning rate, all features) were the best-performing configurations based on average accuracy across five-fold cross-validation. At each depth, bar widths represent softmax-derived scores for each unit. This value is interpreted as the model's relative support for each class. The predicted class corresponds to the unit with the highest score at each depth. This comparison illustrates how architectural and input differences influence class support distributions. In the figure, all formal 'NU' (Upper North Sea Group) formation prefixes have been omitted for clarity. Lithology notation: GY: Gyttja, K: Clay, L: Loam, V: Peat, ZMFO: Sand Median class moderately fine, ZMGO: Sand Median class moderately coarse, ZUGO: Sand Median class extremely coarse, ZZFO: Sand Median class extremely fine, ZZGO: Sand Median class very coarse.

accuracy and Cohen's kappa). The RF1 and RF3 models show more significant differences in the F1 scores, indicating more balanced predictions across under-represented classes. This is supported by the weighted F1-score, showing that model differences are less pronounced after accounting for class distribution. For models using only lithological features (Set 2), the RF2 model consistently underperforms across all metrics by a large margin. In contrast, the NN2 model achieves more consistent results, though it underperforms models using location features (Set 1) or all features (Set 3). Overall, differences in traditional metrics are noticeable but not substantial enough to favour one model type universally, except for F1-scores, where RF1 and RF3 achieve better values.

In contrast, sequence metrics reveal more pronounced differences among models. These metrics evaluate predictions as sequences of units, with some models producing results that more closely resemble the expected order of geological units. As a result, models with similar values for traditional metrics show distinct differences in their sequential performance. For example, the RF3 model, which is the best-performing model using traditional metrics, ranks fifth in both the Sequence Alignment Score (SAS) and the Transition Validation Score (TVS). Notably, the differences between the RF1 and RF3 models are more pronounced in the TVS, resulting in a proportion of 0.66 ± 0.03 of predicted transitions matching known stratigraphic relationships for the RF3 model. In contrast, the RF1 model achieves a value of 0.92 ± 0.01. Similarly, the NN1 and NN3 outperform the RF3 model in all sequence metrics, with the NN1 being consistently the best-performing model across these metrics.

These differences are further illustrated in Fig. 7, which compares classification and transition matrix outputs for the NN3 (64 LSTM units, one-head multi-head attention layer, learning rate = 0.001) and RF3 (mtry = 60) models. Both predictions achieve similar accuracy on a single test fold (0.85), and their class-level prediction patterns are broadly comparable. RF3 outperforms NN3 on the rarest, shallowest units (<1% of the dataset), which NN3 mostly misses. However, this similarity in classification outcomes contrasts with the more apparent distinction shown by the transition matrices, which summarise predicted stratigraphic transitions and compare them to the plausible transitions (i.e. established geological knowledge) defined for the TVS. While both models show similar per-class accuracy for the Breda & Ville Formations (0.78), RF3 frequently places younger units beneath it (e.g. Oosterhout, Inden, or Kiezeloölite formations), even though the Breda & Ville Formations are the expected base of the sequence at this depth range. Misclassifications between the Breda & Ville Formations and the Oosterhout formation, as with many other transitions summarised in the transition matrix, likely reflect lithological variability at their contact, where the lower boundary of Oosterhout Formation is

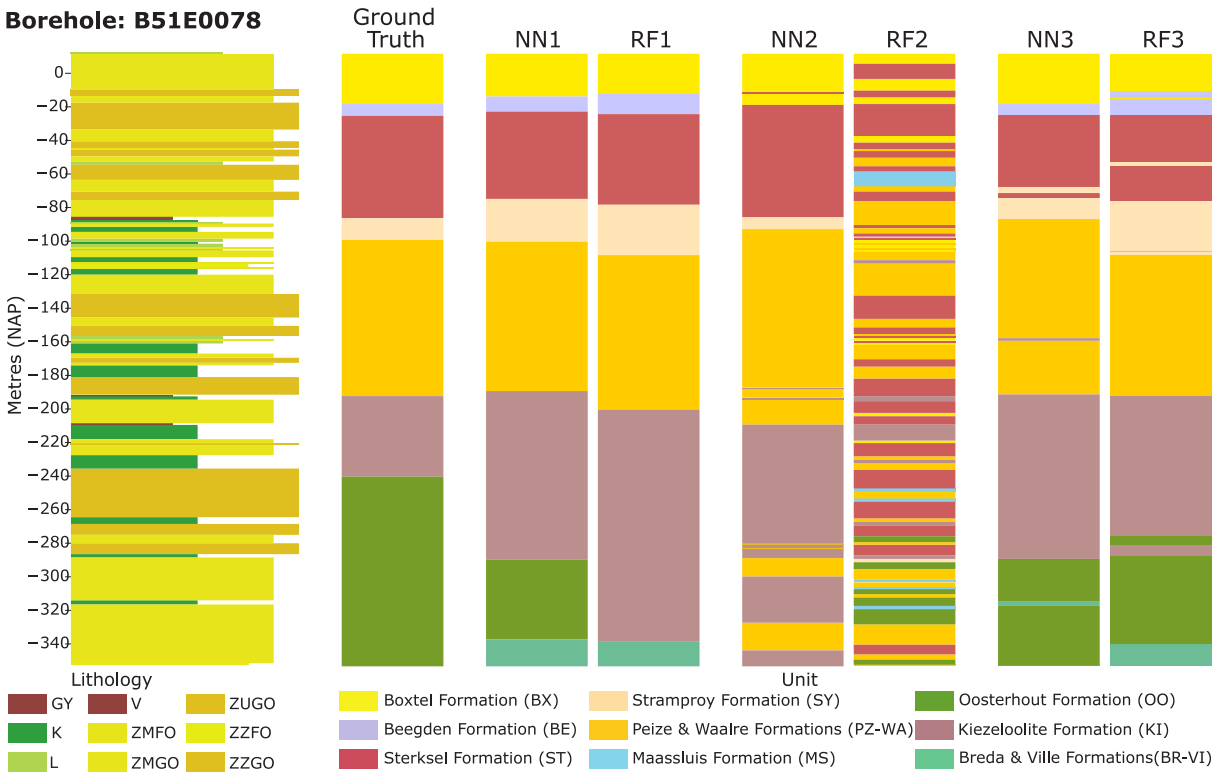

Fig. 5. Lithostratigraphic interpretation of borehole B51E0078. This figure compares the ground-truth interpretation (from the test set of one fold out of five in the stratified group cross-validation) with predictions from six models. Each model uses its best hyperparameter configuration, based on the highest average accuracy across all five folds. In the figure, all formal 'NU' (Upper North Sea Group) formation prefixes have been omitted for clarity. Lithology notation: GY: Gyttja, K: Clay, L: Loam, V: Peat, ZMFO: Sand Median class moderately fine, ZMGO: Sand Median class moderately coarse, ZUGO: Sand Median class extremely coarse, ZZFO: Sand Median class extremely fine, ZZGO: Sand Median class very coarse.

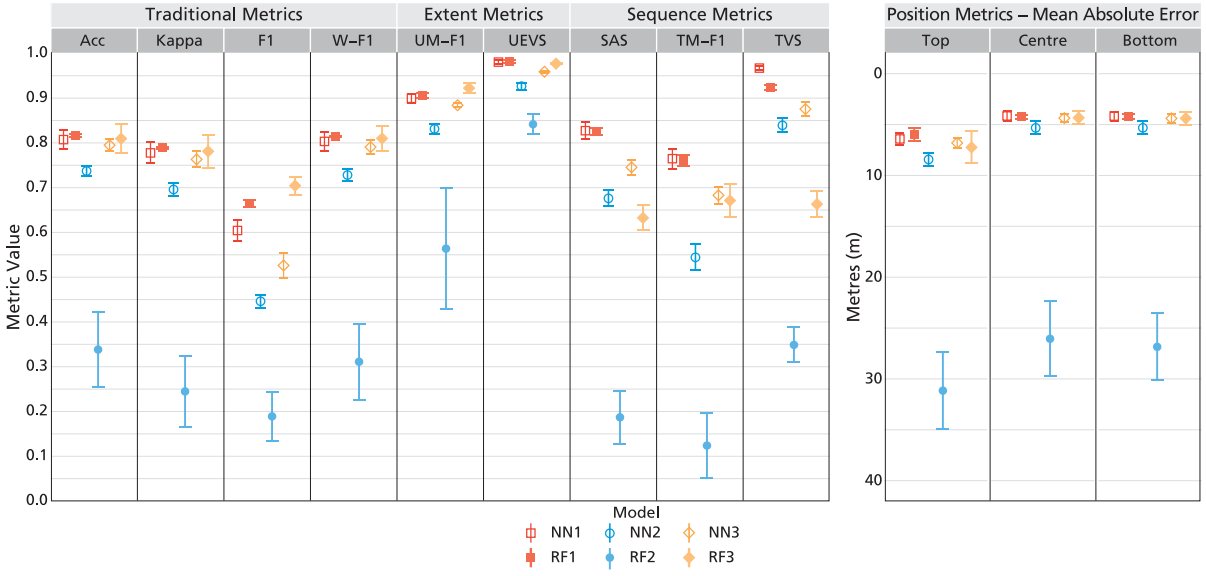

Fig. 6. Model comparison across metrics. The selected models correspond to the hyperparameter configuration with the best metric value per model and feature set. Each value represents the mean across five-fold cross-validation. Error bars indicate one standard deviation above and below the mean. Acc: Accuracy, Kappa: Cohen's Kappa, F1: Macro F1-Score, W-F1: Weighted F1-Score, UM-F1: Unit Match F1-Score, UEVS: Unit Extent Validation Score, SAS: Sequence Alignment Score, TM-F1: Transition Match F1-Score, TVS: Transition Validation Score, MAE-Top: Mean Absolute Error - Top, MAE-Centre: Mean Absolute Error - Centre, MAE-Bottom: Mean Absolute Error - Bottom.

more gradual when its base is sandier (TNO-GDN, 2020). These implausible transitions are less common in the NN3 predictions, highlighting how sequence metrics capture structural inconsistencies that traditional metrics overlook.

Extent and Position metrics show similar results across all models except the RF2 model. On the one hand, extent metrics show a similar pattern as the accuracy and Cohen's kappa metrics, with the RF1 and RF3 models narrowly outperforming the other models. On the other

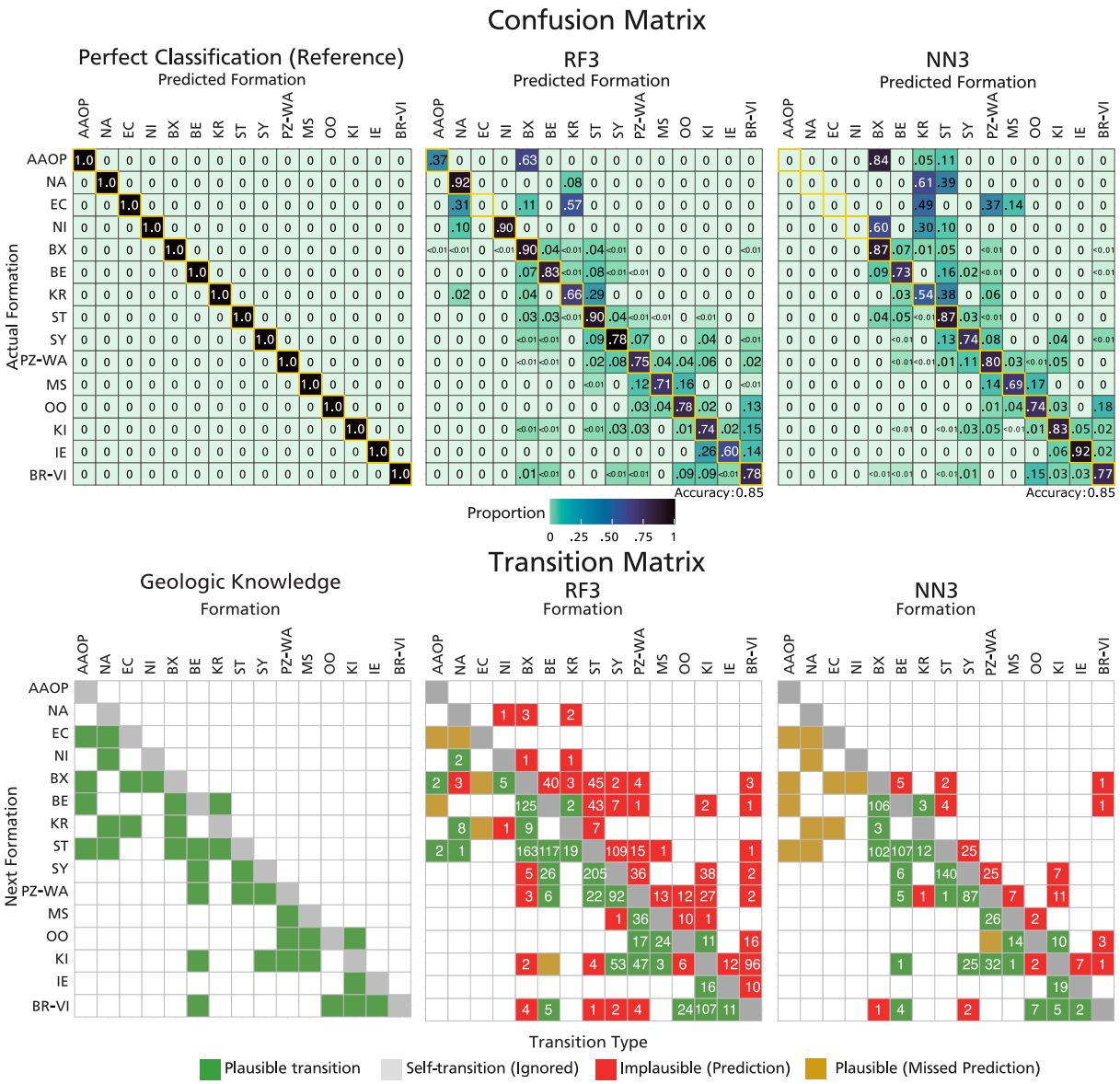

Fig. 7. Confusion matrices (top row) and transition matrices (bottom row) for the ideal case (left), Random Forest (RF3, centre), and Neural Network (NN3, right) predictions on the test set of a representative fold from five-fold cross-validation. Confusion matrices show classification accuracy per formation, with the ideal case displaying a perfect diagonal. Transition matrices illustrate stratigraphic transitions: the ideal case reflects expected transitions based on established geological knowledge (left), compared with RF3 (centre) and NN3 (right) predictions. The *x*-axis represents the 'current' formation, and the *y*-axis the 'next' formation downward in the borehole (i.e. stratigraphically from top to bottom). NN3 predictions more closely align with geologically plausible transitions, exhibiting fewer unlikely contacts (e.g. the Breda and Ville Formations directly above the Boxtel Formation). RF3 (mtry = 60) and NN3 (64 LSTM units, one head in the attention layer, learning rate 0.001) represent the best-performing configurations based on the highest Transition validation score (TVS). In the figure, all formal 'NU' (Upper North Sea Group) formation prefixes have been omitted for clarity.

hand, position metrics show that all models achieve similar results in determining the top, centre, and bottom of a unit except the RF2 model. Thus, the predicted unit position error is around 5 m for the centre and bottom of a unit and 7.5 m for the top. The NN2 model, which lacks location features, performs similarly to other models in position metrics, suggesting that the LSTM units and attention layers capture spatial information through sequence length.

3.2. Results by hyperparameter configuration

Random forest

Model evaluation across hyperparameter configurations shows that, for all RF models, tuning hyperparameters can improve both traditional and geology-informed metrics. The degree of improvement varies by metric and feature set (Fig. 8). Generally, for each feature set, all metrics respond similarly to changes in the mtry hyperparameter, with best results achieved using the same or similar values.

For the RF1 configuration (location features only), traditional metrics indicate the best results with an mtry value of 1. This setup yields up to 0.82 ± 0.005 accuracy and 0.67 ± 0.01 F1-score. Position metrics also favour an mtry value of 1, though other values fall within the standard deviation. In contrast, sequence metrics improve slightly with an mtry of 2. Extent metrics perform best with an mtry of 1 or 2, depending on whether comparisons are made to the ground truth (UM-F1) or external data (UEVS).

The RF2 configuration (lithological features only) consistently performs worse across all metrics. Traditional metrics peak with an mtry value of 10, achieving an accuracy of up to 0.34 ± 0.09 and an F1-score

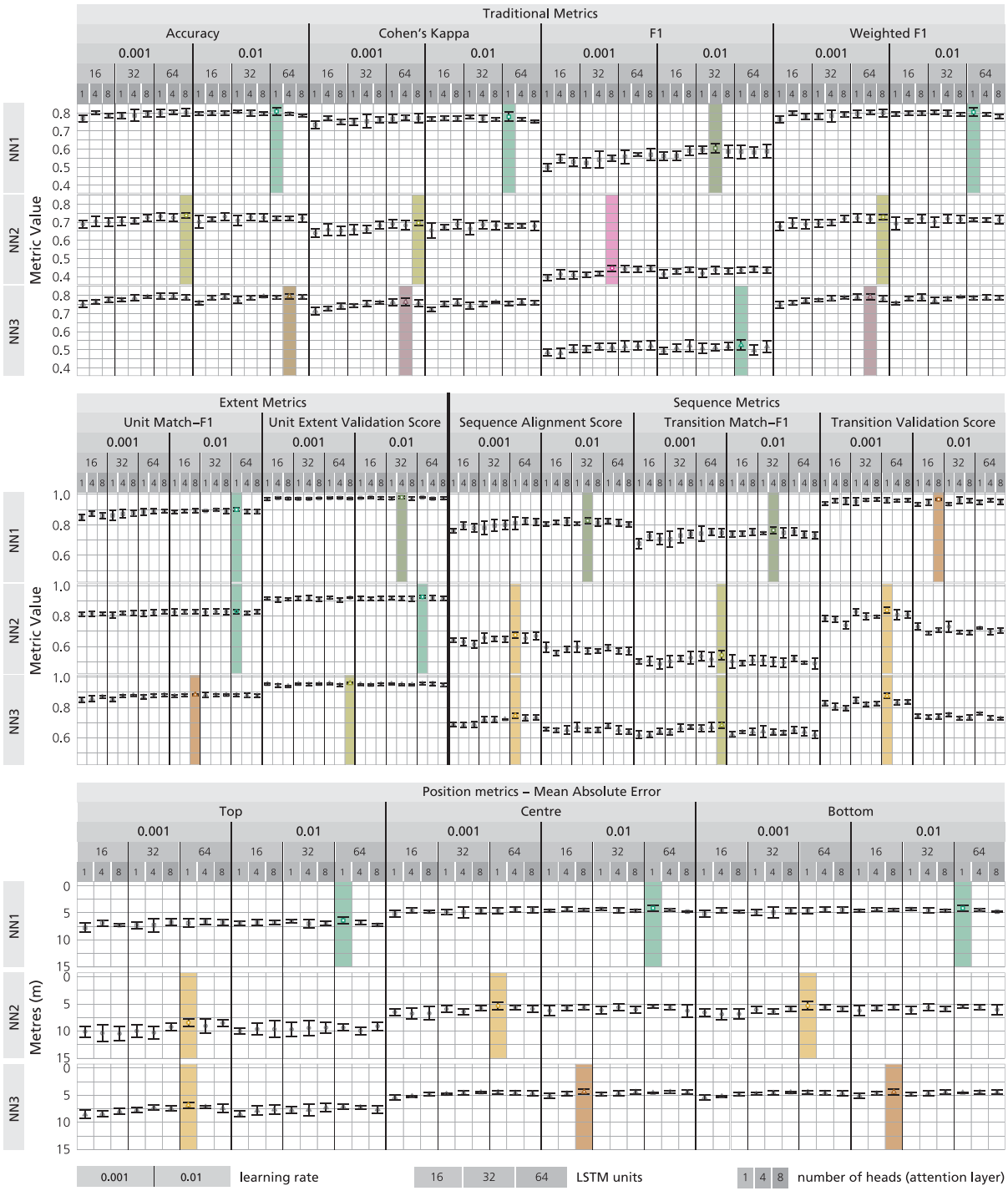

Fig. 8. Model performance for all Random Forest models by hyperparameter configuration. Each value represents the mean across five-fold cross-validation. Error bars indicate one standard deviation above and below the mean. The best hyperparameter configuration per metric is highlighted.

of 0.19 ± 0.06. Adjusting this hyperparameter has little effect here, as most metrics show worse or unchanged performance with increasing values.

In RF3 (combining all features), the optimal mtry value depends on the evaluation metric. Higher mtry values generally lead to better results. Traditional metrics perform best with an mtry value of 40 or 50, while sequence metrics peak at a value of 50 or 60. The optimal RF3 configuration achieves an accuracy of up to 0.81 ± 0.03 and an F1-score of 0.70 ± 0.02. Extent metrics vary more, with optimal results at an mtry value of 30 or 50, depending on whether the comparison is to the ground truth (UM-F1) or external sources (UEVS). Overall, most metrics improve with higher mtry, especially between 40 and 50.

Neural networks

Hyperparameter tuning for the NN models (NN1, NN2, NN3) yields apparent performance differences between best and worst configurations per metric (Fig. 9). However, similar results across settings make it difficult to pinpoint a single best configuration. For NN2 and NN3, the effects of hyperparameters are more pronounced, with better performance associated with a higher number of LSTM units, a lower learning

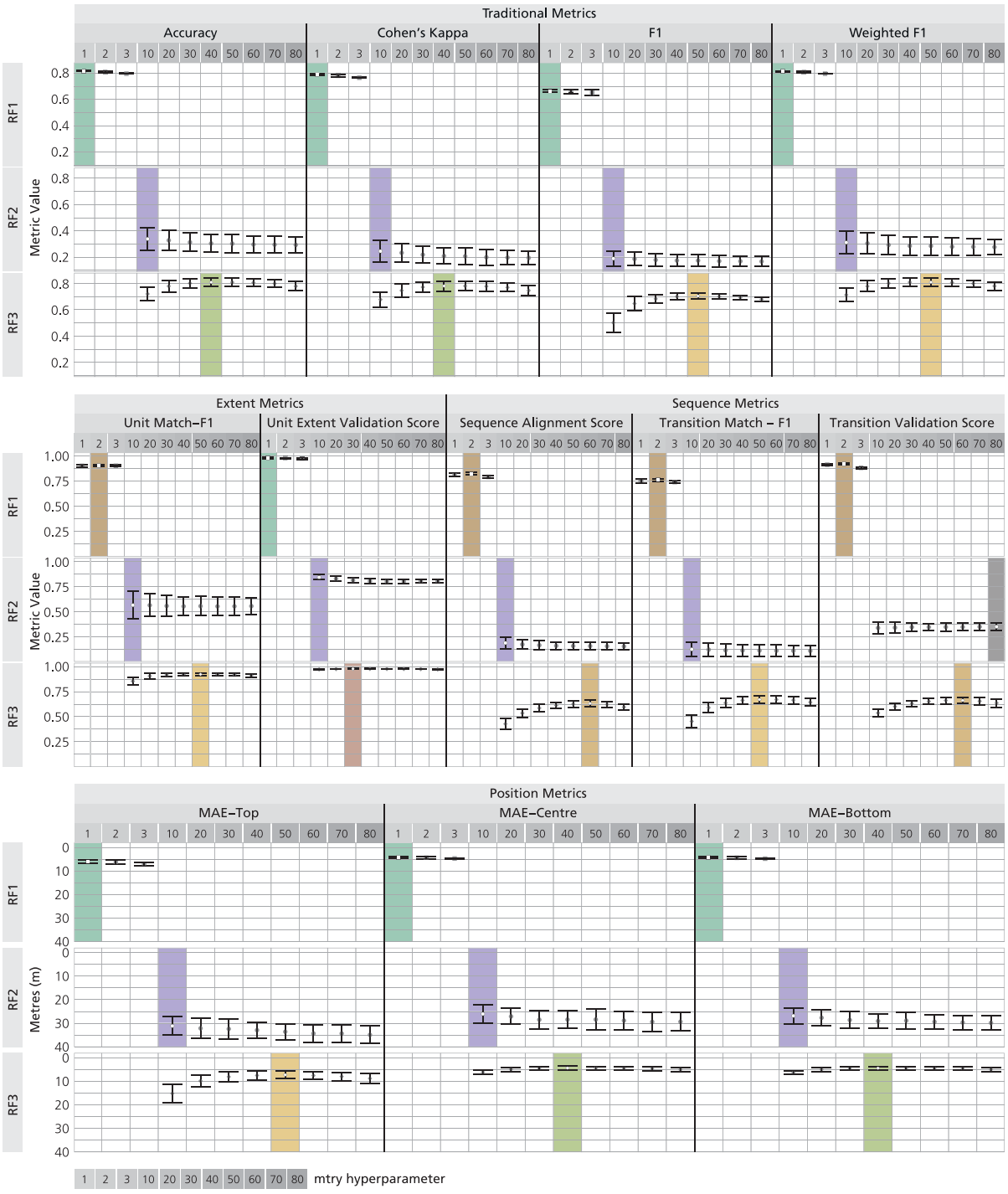

Fig. 9. Model performance for all Neural Network models by hyperparameter configuration. Each value represents the mean across five-fold cross-validation. Error bars indicate one standard deviation above and below the mean. The best hyperparameter configuration per metric is highlighted.

rate, and more heads in the multi-head attention layer, particularly for sequence metrics.

For the NN1 configuration, there is minimal variation across metrics, making it more challenging to determine the optimal setup. Traditional metrics are similar across configurations, with peak values of accuracy (0.81 ± 0.02), Cohen's kappa (0.78 ± 0.03), and weighted F1 (0.8 ± 0.02) achieved using a learning rate of 0.01, 64 LSTM units, and one attention head. This suggests limited sensitivity to hyperparameters, except the F1-Score, which shows a preference for a learning rate of 0.01. Geology-informed metrics show a similar pattern, with modest gains from a 0.01 learning rate and minimal impact from LSTM units or number of attention heads.

The NN2 configuration exhibits more apparent performance shifts across configurations. The best accuracy (0.74 ± 0.01) comes from a learning rate of 0.001, 64 LSTM units, and eight heads, which is also optimal for most other traditional metrics (except overall F1, where differences are minor). In general, increasing the number of LSTM units and attention heads yields slight improvements. Geology-informed metrics (e.g. SAS, TVS) are more sensitive, especially to the learning rate, favouring a value of 0.001. Sequence and position metrics consistently perform best with a learning rate of 0.001, 64 LSTM units,

and one head. These results suggest benefits from a more complex architecture (i.e. more LSTM units) and a lower learning rate.

For the NN3 configuration, results vary more by metric. The best accuracy (0.79 ± 0.01) is achieved with a learning rate of 0.01, 64 LSTM units, and four heads in the attention layer. Increasing the number of LSTM units and the number of heads in the attention layer improves performance across traditional metrics. For geology-informed metrics, extent and position scores are stable across settings, but sequence metrics strongly favour a learning rate of 0.001.

4. Discussion

This study demonstrates how geology-informed metrics enhance model evaluation for predicting lithostratigraphic units. These metrics integrate geological principles and reflect how geologists would evaluate borehole interpretations. Specifically, geology-informed metrics help distinguish models that produce geologically plausible results from those that excel only under traditional evaluation. We demonstrate that traditional and geology-informed metrics yield different best-performing models, underscoring the importance of evaluation metrics. In this section, we discuss how these metrics reflect the strengths and limitations of different models, evaluate the implications for geological modelling and identify potential avenues for future research.

The traditional and geology-informed metrics proposed in this study offer valuable insights into the predictions made by Random Forest (RF) and Neural Network (NN) models. While traditional metrics show that both models can achieve similar classification performance, the geology-informed sequence metrics reveal the superiority of the NN in producing predictions that align more closely with geological principles. Neural Networks can learn complex relationships between features, resulting in fewer transitions between units and more connected units, consistent with geological interpretations. For instance, although similar unique units are predicted per borehole across models (as indicated by the Unit Match - F1 score), NN models distribute these units more consistently, adhering more closely to geological principles. This suggests that the proposed metrics -especially the sequence metrics- provide a more nuanced evaluation of model predictions by incorporating the sequential nature of lithostratigraphic units, which traditional metrics overlook.

Feature selection also affects model performance with metrics favouring different feature sets. For example, models using all features (RF3, NN3) generally achieve higher scores on traditional metrics. In contrast, models with lithological features (RF2, NN2) tend to underperform across most metrics. For sequence metrics, models using only location features (RF1, NN1) often outperform those incorporating all features (RF3, NN3). This strong performance of the position-only models across the sequence metrics suggests that spatial location alone can provide substantial predictive power in our study area, where stratigraphic relationships are relatively consistent and laterally continuous (Fig. 1 B & D). This pattern is illustrated by comparing predicted class voting proportions by depth (Fig. 4), showing how model behaviour varies with different input features. For NN1, the voting proportion changes smoothly with depth, reflecting a gradual interpolation between known stratigraphic positions. In contrast, NN3 exhibits sharper transitions in voting proportions, suggesting that including lithological features enables the model to respond more strongly to local variations in the input features (e.g. mean sand size). While this may increase stratigraphic errors, it may also reflect greater confidence in stratigraphic boundary positions compared to the smoother transitions of the spatial-only model. In addition to the relatively simple geological complexity of the area, the strong class imbalance likely reinforces this effect, as the most frequent units (e.g. Boxtel, Sterksel, Peize & Waalre, and Kiezeloölite formations) can already be predicted with high confidence based on spatial trends alone (Fig. 1. B & D). In areas with significant lateral facies changes, unconformities, or structural deformation, we expect spatial location to become a weaker predictor of stratigraphy, thereby reducing the geological plausibility of interpolation-based predictions.

The results also reveal that NN and RF models integrate sequential information differently into the prediction of lithostratigraphic units. For instance, the NN2 model can capture complex relationships between lithological features without directly relying on location features, significantly outperforming RF2 across all metrics and achieving similar, yet lower, results than other models that incorporate location information. These results demonstrate that a Neural Network with components designed to process sequential data (e.g. Long-Short-Term Memory) can more effectively incorporate lithological data to produce plausible geological predictions than a Random Forest model.

Implications for geological modelling

Machine learning models for interpreting borehole data provide an alternative to probabilistic approaches, such as Markov chain models (e.g. Yin et al., 2022; Eidsvik et al., 2004), which explicitly encode stratigraphic transitions through transition matrices. These matrices reflect assumptions about probability distributions derived from observed data and incorporate additional constraints such as stationarity and fixed-order dependencies. In contrast, ML models learn complex, potentially non-linear and high-dimensional relationship patterns directly from the data without prior assumptions (Qi and Carr, 2006). A common limitation of non-probabilistic approaches, including many ML models (e.g. Tokpanov et al., 2020; Wedge et al., 2019), is that they may produce geologically implausible outputs (e.g. implausible transitions between units) if the model does not incorporate geological context. In our case study, where strict stratigraphic rules constrain the vertical order of lithostratigraphic units, our Neural Network uses information from neighbouring depths to better capture this structure, resulting in predictions that are often geologically plausible, even without explicit post-processing. While probabilistic approaches have the advantage of encoding plausible stratigraphic transitions, our domain-oriented evaluation of borehole labelling suggests that well-designed ML models — particularly Neural Networks with appropriate architecture, set of features, and hyperparameters — can achieve high accuracy while also partially capturing stratigraphic relationships.

The varying performance across geology-informed metrics for models that appear similar based on traditional metrics suggests that fundamental aspects of subsurface structure might be overlooked during model evaluation. Although a post-processing step can correct some interpretation errors in the case of lithostratigraphic units, other less restrictive yet sequential problems, such as lithofacies predictions, lack obvious post-processing solutions. As a result, minor performance differences between models can translate into significant changes in the predicted subsurface structure. For instance, hydrogeological models are sensitive to the distribution and connectivity of different units (e.g. sandy versus shaly sediment sequences), which, as shown in this study, vary across seemingly similar models based on traditional metrics. Therefore, geology-informed metrics can help identify models more likely to produce geologically plausible predictions.

For nationwide 3D subsurface models, such as the Digital Geological Model (DGM) (Gunnink et al., 2013), which uses interpreted boreholes as input for spatial interpolation, geology-informed metrics can help identify predictions aligned with known geological characteristics of the area. The TNO–Geological Survey of the Netherlands manages a dataset of over 600 000 boreholes. However, only a small subset (5%) is typically used for the DGM, illustrating the vast scale of potential data available. With many boreholes lacking expert interpretations, automated and semi-automated methods that minimise manual corrections are crucial to increasing the number of usable boreholes for three-dimensional geological modelling. Therefore, automated workflows should prioritise adherence to known geological relationships in the area (e.g. stratigraphic relationships) over maximising classification performance using traditional metrics. The metrics proposed in this

study can help identify prediction models, feature sets, and hyperparameter configurations that produce outputs more closely aligned with geological principles, thereby reducing the need for post-processing tasks.

Although beyond the scope of this study, our results suggest using domain-specific metrics as loss functions in model training to enhance performance by optimising the geological plausibility of predictions. Despite the ability of Neural Networks to learn sequential relationships, our experiments show that even models outperforming on sequence metrics do not fully capture the relatively simple stratigraphic relationships of the area. Standard convex loss functions, such as the categorical cross-entropy used in our case study, may not be optimal given the inherent interpretative uncertainty of stratigraphic labels and the importance of sequential relationships. This results in predictions that incorporate spatial relationships not observed in the training data and that would be immediately flagged as geologically implausible by an expert. Therefore, one promising direction is the use of loss functions tailored to geological interpretation tasks. For instance, Hillier et al. (2023) apply this idea to a three-dimensional interpolation task, showing that stratigraphic consistency can be enforced during model training, which could be adapted for automated borehole interpretation. In parallel, robust loss functions have been developed to handle categorical label noise, such as those based on smooth non-convex formulations for large-margin classification (Feng et al., 2016), providing a complementary strategy to address the label noise and uncertainty inherent in stratigraphic interpretation data. Together, these approaches suggest that better-aligned loss functions could improve the geological plausibility of automated predictions.

Limitations

Implementing the proposed geology-informed metrics for evaluating automated lithostratigraphic interpretations of boreholes tested in this work has several limitations.

First and foremost, most metrics tested in this study rely on the assumption that the provided labels represent a ground truth. However, lithostratigraphic labels are based on expert interpretations and have an inherent uncertainty that is not systematically quantified. The interpretative nature of the task introduces label noise and imposes a performance ceiling on model evaluation. For instance, the interpretative variability of experts defining formation boundaries has been quantified in a limited number of studies, which report errors in interpreted boundary positions in cross-section and borehole experiments in the UK ranging from ± 7 to ± 18 m across different sites, with standard deviations between 2.7 and 6.0 m depending on geological context and interpreter-specific factors (Randle et al., 2018; Lark et al., 2014). This implies automated interpretations may already fall within the expected expert variability. However, metrics relying on direct ground-truth comparisons still penalise any deviation from reference labels.

While interpretative variability influences the assumptions underlying most evaluation metrics, we expect geologists to agree on broader geological concepts, such as those tested in the proposed geology-informed metrics. For example, experts may differ on the precise location of certain stratigraphic boundaries, but interpretations are expected to remain consistent with an established stratigraphic framework. Therefore, any ML-based prediction should be evaluated not only against individual labels but also against this framework. In this context, metrics such as the Transition Validation Score (TVS) are expected to be less sensitive to label noise and to offer a higher performance ceiling. Although we did not explicitly estimate the performance ceiling of these metrics in this study area, we consider that the proposed metrics, particularly those that do not incorporate ground-truth comparisons, offer valuable tools for distinguishing models that produce plausible interpretations. Formal quantification of the performance ceiling represents an important avenue for future research.

Second, the case study focuses on the Roer Valley Graben, characterised by relatively simple stratigraphic relationships that are not generalisable to other geological settings. Therefore, the metrics' ability to distinguish plausible geological predictions might not transfer directly to more complex contexts. For example, sequential metrics such as the Transition Validation Score, which incorporates external geological information (Fig. 7), may be less restrictive in other geological settings. In our scenario, 15 stratigraphic units define 225 possible transitions between lithostratigraphic units, but only 50 of these have been observed. While more complex geological areas, such as faulted or tilted regions or areas with varying degrees of erosive contacts, would likely contain additional plausible transitions, we expect the set of implausible transitions to remain significant, allowing sequence metrics to remain effective. These assumptions may not hold for other classification targets where ordering is less constrained by established sequences (e.g. lithologies or lithofacies). Testing the proposed metrics in other geological settings, particularly across basins, would offer valuable insights.

Third, while our study addressed aspects of model uncertainty via five-fold cross-validation, capturing uncertainty related to data sampling and model variability, it does not quantify the geological uncertainty of the predictions. Geological uncertainty is addressed in some probabilistic approaches for labelling borehole data using Bayesian methods, which provide posterior probability distributions reflecting uncertainty in stratigraphic interpretations (e.g. Yin et al., 2022; Eidsvik et al., 2004). However, this aspect is not captured by our current evaluation. Although Neural Network predictions can incorporate uncertainty estimation methods to capture predictive uncertainty (e.g. Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017), our geology-informed metrics evaluate predictions based on the most likely lithostratigraphic class at each depth. Therefore, these metrics assess model performance using class assignments rather than incorporating the full probabilistic distribution of predictions. Incorporating explicit geological uncertainty quantification into Neural network models remains a crucial direction for future research.

Fourth, the proposed metrics have inherent limitations in evaluating predictions of lithostratigraphic units using borehole information. For instance, position metrics cannot be computed if a unit is absent in the prediction or ground truth. Similarly, position metrics are sensitive to noise in predictions and the selected criteria to define a position, affecting the detection of a unit's top, centre, and bottom. Extent metrics such as the Unit Extent Validation Score also have limitations as they might only be informative for large scales where geological units have distinct spatial distribution patterns. In contrast, sequence metrics are only relevant for tasks where order is critical, such as lithostratigraphic predictions, but not for other problems like facies or physical property predictions.

Lastly, this study did not perform feature selection, which is a standard step when dealing with large numbers of input variables. Instead, we intentionally evaluated model performance using predefined, geologically meaningful feature sets: location-only (set 1), lithological-only (set 2), and a combination of both (set 3), to compare how models respond to different types of input information. For example, while location-only features provided reasonable predictions, lithological features showed promising results with the Neural Network (NN2) versus the RF models. This approach allowed us to assess the relative importance of spatial versus lithological information rather than optimising feature subsets through statistical methods. Nonetheless, the geology-informed metrics developed here provide a valuable foundation for future workflows that could incorporate geological intuition into formal feature selection, potentially identifying the most critical features improving predictions.

Despite the limitations, the geology-informed metrics proposed in this study provide valuable insights into model performance by emphasising geological aspects that traditional metrics may overlook. Although further work is needed to assess their generalisability to other geological settings and prediction tasks, this use case illustrates the potential of these metrics for evaluating the geological plausibility of ML-generated borehole interpretations.

5. Conclusion

In this study, we presented a set of geology-informed metrics to evaluate the performance of automated prediction models for lithostratigraphic borehole interpretation. To illustrate their usefulness, we applied the metrics to two distinct model types — Random Forests and Neural Networks — using a case study in the Roer Valley Graben. The results show that geology-informed metrics, particularly sequence metrics, capture significant differences among models that traditional metrics overlook. While Neural Networks are expected to excel in sequential tasks, our case study shows that differences emerge only when evaluated with metrics aligned to the task's sequential nature. The proposed metrics provide an informative and complementary perspective to traditional metrics, ultimately enabling us to quantify the geological plausibility of model predictions.

Our findings also underscore the value of domain-specific metrics to reveal performance advantages. In our case study, adhering to stratigraphic order is a fundamental aspect of the prediction task. Incorporating these metrics during training could help geologists identify models that produce more geologically plausible predictions. This approach reduces the need for post-processing, simplifying automated borehole interpretation. Ultimately, this would increase the number of borehole interpretations integrated into three-dimensional subsurface models. Future work will extend the application of these metrics in other sedimentary basins and explore their integration as loss functions in Machine learning models to further improve the interpretation of lithostratigraphic units from borehole data.

CRediT authorship contribution statement

Sebastián Garzón: Writing – original draft, Visualization, Software, Methodology, Investigation, Conceptualization. **Willem Dabekaussen:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Freek S. Busschers:** Writing – review & editing, Supervision, Conceptualization. **Eva De Boever:** Writing – review & editing, Supervision, Conceptualization. **Siamak Mehrkanoon:** Writing – review & editing, Supervision, Conceptualization. **Derek Karssenberg:** Writing – review & editing, Supervision, Conceptualization.

Code availability section

Library name: geology-informed-borehole-metrics
License: MIT LICENSE
Contact: j.s.garzonalvarado@uu.nl
Hardware requirements: Standard systems suffice
Program language: R
Software required: R & Python
Program size: 25.1 MB
The source codes are available for downloading at the link:
https://github.com/SbastianGarzon/geology-informed-borehole-metrics

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge the financial and logistical support provided by TNO – Geological Survey of the Netherlands. We also thank Jan Stafleu for his assistance in facilitating the use and interpretation of the datasets employed in this study.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cageo.2025.106043.

Data availability

The data analysed in this study is available via Zenodo: https://www.doi.org/10.5281/zenodo.14859951.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. URL: https://www.tensorflow.org/. Software available from tensorflow.org.

Bhattacharya, S., 2021. Summarized applications of machine learning in subsurface geosciences. In: A Primer on Machine Learning in Subsurface Geosciences. Springer International Publishing, Cham, pp. 123–165. http://dx.doi.org/10.1007/978-3-030-71768-1_5, Chapter 5.

Bosch, J., 2000. Standaard Boorbeschrijvingsmethode: Versie 5.1. Nederlands Instituut voor Toegepaste Geowetenschappen TNO.

Botchkarev, A., 2019. A new typology design of performance metrics to measure errors in machine learning regression algorithms. Interdiscip. J. Inf. Knowl. Manag. 14, 45–79. http://dx.doi.org/10.28945/4184.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Eidsvik, J., Mukerji, T., Switzer, P., 2004. Estimation of geological attributes from a well log: An application of hidden Markov chains. Math. Geol. 36 (3), 379–397. http://dx.doi.org/10.1023/B:MATG.0000028443.75501.d9.

Feng, Y., Yang, Y., Huang, X., Mehrkanoon, S., Suykens, J.A., 2016. Robust support vector machines for classification with nonconvex and smooth losses. Neural Comput. 28 (6), 1217–1247.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. 15 (90), 3133–3181, URL: http://jmlr.org/papers/v15/delgado14a.html.

Fullagar, P.K., Zhou, B., Biggs, M., 2004. Stratigraphically consistent autointerpretation of borehole data. J. Appl. Geophys. 55 (1), 91–104. http://dx.doi.org/10.1016/j.jappgeo.2003.06.010, URL: https://www.sciencedirect.com/science/article/pii/S0926985103000727. Non-Petroleum Applications of Borehole Geophysics.

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Balcan, M.F., Weinberger, K.Q. (Eds.), Proceedings of the 33rd International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 48, PMLR, New York, New York, USA, pp. 1050–1059, URL: https://proceedings.mlr.press/v48/gal16.html.

Geluk, M., Duin, E.T., Dusar, M., Rijkers, R., Van den Berg, M., Van Rooijen, P., 1995. Stratigraphy and tectonics of the Roer Valley Graben. Geol. Mijnb. 73, 129.

v. Gessel, S., Hintersberger, E., v. Ede, R., ten Veen, J., Doornenbal, H., Diepolder, G.W., den Dulk, M., Hamiti, S., Vukzaj, N., Çako, R., Prendi, E., Ceroni, M., Mara, A., Barros, R., Tovar, A., Britze, P., Baudin, T., Stück, H., Jähne-Klingberg, F., Jahnke, C., Höding, T., Malz, A., Kristjánsdóttir, S., Þorbergsson, A., Di Manna, P., D'Ambrogi, C., Congi, M., Lazauskienė, J., Andriuškevičienė, G., Baliukevičius, A., Jarosiński, M., Gogołek, T., Stępień, U., Krzemińska, E., Salwa, S., Habryn, R., Aleksandrowski, P., Szynkaruk, E., Konieczyńska, M., Ressurreição, R., Machado, S., Moniz, C., Sampaio, J., Dias, R., Carvalho, J., Fernandes, J., Ramalho, E., Filipe, A., Celarc, B., Atanackov, J., Jamšek Rupnik, P., Shevchenko, A., Melnyk, I., Lapshyna, A., 2021. The HIKE European fault database (EFDB) compiled in the framework of the GeoERA project HIKE (2018–2021). https://egdi.geology.cz/record/basic/5edf7bd4-9270-4188-b69d-7ddd0a010833. (Accessed 26 May 2025).

Grandini, M., Bagli, E., Visani, G., 2020. Metrics for multi-class classification: An overview. URL: https://arxiv.org/abs/2008.05756, arXiv:2008.05756.

Gunnink, J., Maljers, D., van Gessel, S., Menkovic, A., Hummelman, H., 2013. Digital geological model (DGM): A 3D raster model of the subsurface of the netherlands. Neth. J. Geosci. - Geol. Mijnb. 92 (1), 33–46. http://dx.doi.org/10.1017/S0016774600000263.

Hillier, M., Wellmann, F., de Kemp, E.A., Brodaric, B., Schetselaar, E., Bédard, K., 2023. Geoinr 1.0: An implicit neural network approach to three-dimensional geological modelling. Geosci. Model. Dev. 16 (23), 6987–7012. http://dx.doi.org/10.5194/gmd-16-6987-2023, URL: https://gmd.copernicus.org/articles/16/6987/2023/.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Hummelman, J., Maljers, D., Menkovic, A., Reindersma, R., Stafleu, J., Vernes, R., 2019. Totstandkomingsrapport Digitaal Geologisch Model (DGM). Report: TNO 11653.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. URL: https://arxiv.org/abs/1612.01474, arXiv:1612.01474.

Lark, R.M., Thorpe, S., Kessler, H., Mathers, S.J., 2014. Interpretative modelling of a geological cross section from boreholes: Sources of uncertainty and their quantification. Solid Earth 5 (2), 1189–1203. http://dx.doi.org/10.5194/se-5-1189-2014, URL: https://se.copernicus.org/articles/5/1189/2014/.

Loo, M.P.v.d., 2014. The stringdist package for approximate string matching. R J. 6, 111–122, https://rjournal.github.io/.

Naidu, G., Zuva, T., Sibanda, E.M., 2023. A review of evaluation metrics in machine learning algorithms. In: Silhavy, R., Silhavy, P. (Eds.), Artificial Intelligence Application in Networks and Systems. Springer International Publishing, Cham, pp. 15–25.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Qi, L., Carr, T.R., 2006. Neural network prediction of carbonate lithofacies from well logs, big bow and sand arroyo creek fields, Southwest Kansas. Comput. Geosci. 32 (7), 947–964. http://dx.doi.org/10.1016/j.cageo.2005.10.020, URL: https://www.sciencedirect.com/science/article/pii/S0098300405002396. Computer Simulation of natural phenomena for Hazard Assessment.

Randle, C.H., Bond, C.E., Lark, R.M., Monaghan, A.A., 2018. Can uncertainty in geological cross-section interpretations be quantified and predicted? Geosphere 14 (3), 1087–1100. http://dx.doi.org/10.1130/GES01510.1, arXiv:https://pubs.geoscienceworld.org/gsa/geosphere/article-pdf/14/3/1087/4224351/1087.pdf.

Stafleu, J., Busschers, F.S., van der Meulen, M.J., den Dulk, M., Gunnink, J.L., Maljers, D., Hummelman, J.H., Schokker, J., Vernes, R.W., Stam, J., Dabekaussen, W., ten Veen, J.H., Doornenbal, H., Kars, R., de Bruijn, R., 2025. Geological subsurface models of the netherlands. In: Geology of the Netherlands: Second Edition. Amsterdam University Press, pp. 849–894, URL: http://www.jstor.org/stable/jj.27435714.28.

Stafleu, J., Maljers, D., Busschers, F., Schokker, J., Gunnink, J., Dambrink, R., 2021. Models created as 3-D cellular voxel arrays. In: Applied Multidimensional Geological Modeling. John Wiley & Sons, Ltd, pp. 247–271. http://dx.doi.org/10.1002/9781119163091.ch11, chapter 11. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119163091.ch11, URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119163091.ch11.

Stafleu, J., Maljers, D., Gunnink, J., Menkovic, A., Busschers, F., 2011. 3D modelling of the shallow subsurface of Zeeland, the Netherlands. Neth. J. Geosci. - Geol. Mijnb. 90 (4), 293–310. http://dx.doi.org/10.1017/S0016774600000597.

Stumpf, A.J., Keefer, D.A., Turner, A.K., 2021. Overview and history of 3-D modeling approaches. In: Applied Multidimensional Geological Modeling. John Wiley & Sons, Ltd, pp. 93–112. http://dx.doi.org/10.1002/9781119163091.ch5, chapter 5. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119163091.ch5, URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119163091.ch5.

Thomason, J.F., Keefer, D.A., 2021. Model creation using stacked surfaces. In: Applied Multidimensional Geological Modeling. John Wiley & Sons, Ltd, pp. 211–233. http://dx.doi.org/10.1002/9781119163091.ch9, chapter 9. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119163091.ch9, URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119163091.ch9.

TNO–GDN, 2014. BRO DGM v2.2. https://www.dinoloket.nl/en/subsurface-models/map. (Accessed 06 February 2024).

TNO-GDN, 2020. Stratigraphic nomenclature of the Netherlands. https://www.dinoloket.nl/en/stratigraphic-nomenclature.(Accessed 06 February 2024).

TNO-GDN, 2023. BRO GeoTOP v1.6. https://www.dinoloket.nl/en/subsurface-models/map. (Accessed 06 February 2024).

TNO-GDN, VITO, GSB, 2014. H3O-Roerdalslenk. https://www.dinoloket.nl/en/downloads-project-h3o-roerdalslenk. Downloaded on 2024-02-06.

TNO-GDN, VITO, GSB, 2017. H3O-De Kempen. https://www.dinoloket.nl/downloads-project-h3o-de-kempen. Downloaded on 2024-02-06.

Tokpanov, Y., Smith, J., Ma, Z., Deng, L., Benhallam, W., Salehi, A., Zhai, X., Darabi, H., Castineira, D., 2020. Deep-learning-based automated stratigraphic correlation. In: Conference, S.A.T., Exhibition (Eds.), SPE Annual Technical Conference and Exhibition. D022S061R020. http://dx.doi.org/10.2118/201459-MS, arXiv:https://onepetro.org/SPEATCE/proceedings-pdf/20ATCE/20ATCE/D022S061R020/3945215/spe-201459-ms.pdf.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Vol. 30, Curran Associates, Inc., pp. 6000–6010, URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wedge, D., Hartley, O., McMickan, A., Green, T., Holden, E.-J., 2019. Machine learning assisted geological interpretation of drillhole data: Examples from the pilbara region, western Australia. Ore Geol. Rev. 114, 103118. http://dx.doi.org/10.1016/j.oregeorev.2019.103118, URL: https://www.sciencedirect.com/science/article/pii/S0169136819302148.

Wright, M.N., Ziegler, A., 2017. ranger: A fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. 77 (1), 1–17. http://dx.doi.org/10.18637/jss.v077.i01.

Yang, Y., Wang, J., Li, Z., Liu, N., Liu, R., Gao, J., Wei, T., 2023. Automated stratigraphic correlation of well logs using attention based dense network. Artif. Intell. Geosci. 4, 128–136. http://dx.doi.org/10.1016/j.aiig.2023.09.001, URL: https://www.sciencedirect.com/science/article/pii/S2666544123000278.

Yin, Z., Amaru, M., Wang, Y., Li, L., Caers, J., 2022. Quantifying uncertainty in downscaling of seismic data to high-resolution 3-D lithological models. IEEE Trans. Geosci. Remote Sens. 60, 1–12. http://dx.doi.org/10.1109/TGRS.2022.3153934.

Zhou, C., Ouyang, J., Ming, W., Zhang, G., Du, Z., Liu, Z., 2019. A stratigraphic prediction method based on machine learning. Appl. Sci. 9 (17), http://dx.doi.org/10.3390/app9173553, URL: https://www.mdpi.com/2076-3417/9/17/3553.