Automated Classification of Estuarine Sub-depositional Environment Using Sediment Texture

J. E. Houghton[1], T. E. Nichols[1], J. Griffiths[1,2], N. Simon[1] J. E. P. Utley[1], R. A. Duller[1], and R. H. Worden[1]

[1]Diagenesis Research Group, Department of Earth, Ocean, and Ecological Sciences, School of Environmental Sciences, University of Liverpool, 4 Brownlow Street, Liverpool, L69 3GP, UK

[2]National Nuclear Laboratory, 5th Floor, Chadwick House, Birchwood Park, Risley, Warrington, WA3 6AE

Corresponding authors: James Houghton (J.E.Houghton@liverpool.ac.uk), Richard Worden (R.Worden@liverpool.ac.uk)

Key Points:

- We propose a machine learning workflow to predict sub-depositional environment in an estuary using sediment texture (e.g., sorting).
- Two surface-calibrated predictive models are presented to automatically classify estuarine core sediment samples.
- Application of the predictive models to core data allows for an unbiased interpretation of the sandy estuarine sequence.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1029/2022JF006891.

Abstract

Interpretation of unconsolidated Quaternary sedimentary core is difficult if key diagnostic features are obscured or not present, therefore traditional facies analysis is challenging. However, sediment texture remains a universal attribute which can be used to interpret sedimentary core. Here we present an automated classification workflow which implements Extreme Gradient Boosting and Bayesian Optimization of hyperparameters to differentiate estuarine sub-depositional environments. We use nineteen textural attributes, measured using laser particle size analysis of surface sediment samples from the Ravenglass Estuary, Cumbria, northwest England, to make unbiased classification of sub-depositional environment and estuarine zone. Two predictive models created using the automated workflow are presented and evaluated using a suite of evaluation metrics, confusion matrices, and spatial analysis to understand their geological implications. Model 1 keeps all sub-depositional environments discrete and has an overall accuracy of 68.96%. Model 2 merges related sub-depositional environments to form inner-coarse and outer-estuary zones and has an overall accuracy of 84.14%. Both models have been applied to textural data obtained at 5cm intervals from a Holocene core drilled through a tidal bar in the Ravenglass estuarine succession, NW England, to classify palaeo sub-depositional environment. Predictive output of the models suggests that the core consistently experienced inner estuary deposition; all inner estuary environments are represented in the core. The workflow presented here could be applied to datasets from other marginal marine depositional systems to enhance the interpretation of their subsurface deposits. Ultimately, detailed interpretations of ancient, buried deposits could be made using models derived from analogous modern systems.

Plain Language Summary

Geoscientists typically rely on characteristic structures when interpreting core. In Quaternary core, distinguishing structures can be absent or obscured as sediment is poorly consolidated (falls apart easily). In this case, a geoscientist must rely on alternative methods that can aid core interpretation. Using proven statistical links between the size, and distribution of sand grains and sedimentary environment, we have

developed a new predictive machine learning model, using freely-available open source software, trained to surface sediment from a well-studied estuary in the UK. The new surface calibrated predictive model can be used to aid a geoscientist with interpretation of core drilled into the estuary to predict how depositional environment changed with time. We have applied the predictive models to a section of core, that lacks characteristic and distinguishing features, and made subtle interpretations that had been missed during traditional sedimentary core interpretation. The use of the new predictive machine learning model permits unbiased interpretations, and should be used alongside established core interpretation methods. The models produced here are flexible and can be adapted for use as part of any classification problem that uses either numerical, or categorical data.

1 Introduction

Machine learning is the use of computation methods, or algorithms, to analyze data, identify patterns, and make accurate predictions to classify data (Mohri et al., 2018). Classification problems within geoscience, such as lithology identification from well log data, seismic interpretation, and classification of remote sensing data, can all benefit from machine learning. Machine learning algorithms are divided into unsupervised and supervised learning algorithms. Unsupervised learning algorithms can identify patterns in data that do not carry a classification label, whereas supervised machine learning approaches, using labelled training data, tends to be more accurate (Mohamed et al., 2019; Sun et al., 2020).

Extreme Gradient Boosting (XGBoost) is a classification tree-based supervised machine learning algorithm suitable for application to classification and regression problems (Chen & Guestrin, 2016). The advantage of XGBoost over other supervised machine algorithms (e.g., random forest and neural networks) is improved optimization and computation, which enhances performance owing to the regularized model that controls overfitting (Chen & Guestrin, 2016). Gradient tree boosting algorithms combine many weak decision tree learners, by continuous residual learning, into one strong learner (Friedman, 2001; Friedman, 2002). XGBoost has been used by data scientists to achieve state-of-the-art

results for many complex classification problems (Zheng et al., 2022; Zhong et al., 2020). Here, we will use XGBoost to understand and classify estuarine sediment.

An estuary is the seaward portion of a drowned valley system where sediment is sourced from both the hinterland, via aeolian and/or fluvial processes, and from offshore, via tidal currents and wave action (Dalrymple et al., 1992). Sediment provenance (source bedrock type and mineralogy) and the biophysiochemical processes that occur in the hinterland and during sediment transport control the bulk grain size distribution of sediment entering an estuary (Blott & Pye, 2001; Visher, 1969). Within an estuary, sediments are hydrodynamically sorted by a mixture of fluvial, tidal and wave processes leading to a distinct set of sub-depositional environments (geomorphic facies) at the surface and facies and facies associations preserved in the subsurface (Boyd et al., 2006; Dalrymple et al., 1992; Heap et al., 2004).

Palaeo-environmental reconstruction using sediment cores must rely on sedimentary characteristics alone if diagnostic sedimentary structures and ichnofabrics are sparse or obscured, which means differentiating sub environments using a facies analysis approach (e.g. Walker, 1990) can be subjective, particularly in sand rich environments. As such, sediment texture (grain size distribution) remains a universal discriminator between sub-depositional environments, which can be more robust if a direct link between the texture of surface sub-depositional environments and subsurface palaeo sub-depositional environment is proven.

There have been numerous attempts to relate sediment texture to depositional environment and the processes within those environments. Early approaches employed statistical measurements of particle size distribution of sand to discriminate between beach, dune, river, and aeolian environments (Biederman, 1962; Friedman, 1961; Mason & Folk, 1958; Moiola & Spencer, 1979; Moiola et al., 1974; Sevon, 1966; Vincent, 1998). More advanced approaches employ Analysis of Variance (ANOVA), bivariate or multivariate discrimination, such as principal component analysis (PCA), with increased success over simple statistical measurements (Flood et al., 2015; Purkait & Das Majumdar, 2014; Simon et al., 2021;

Zheng & Wu, 2021; Zubillaga & Edwards, 2005). Simon et al. (2021) proved a strong statistical link between surface sub-depositional environment and sediment texture in the Ravenglass Estuary, UK, and devised a simple machine learning classification scheme for the classification of sub-depositional environment from sediment texture data.

We build on the work of Simon et al. (2021) by developing a machine learning workflow to test the applicability and validity of the XGBoost algorithm as a predictive tool; the aim was to robustly characterize estuarine sub-depositional environments based solely on sediment texture obtained through Laser Particle Size Analysis (LPSA). Here, we have calibrated predictive XGBoost models using surface sediment samples from the Ravenglass Estuary in order to predict palaeo sub-depositional environments in a modern-day estuarine tidal bar core, from the Ravenglass Estuary. The models have been evaluated using 4-fold cross-validation and a suite of established evaluation metrics suited to multi-class problems. Evaluating the models identified where predictions are strong and where they are weak. The surface calibrated XGBoost models were then applied to 'unseen' subsurface core data in order to predict paleo sub-depositional environment. The application of surface calibrated modes to subsurface core data allowed an unbiased interpretation based on sediment texture and adds an additional tool to core interpretation, especially when key diagnostic features are obscured. The classification of sub-depositional environment in a geotechnical core are compared to a traditional sedimentary logging and facies interpretation adopted by McGhee et al. (2022) in order to test the validity of the predictive models.

2 The Ravenglass Estuary

2.1 Geological setting, Geomorphology, and Estuarine Hydrodynamics

The Ravenglass Estuary is located in Cumbria, northwest England (Figure 1a). The estuary is a macro-tidal (> 7 m tidal range), tide-dominated system that covers an area of 5.6 km^2, of which 86 % is intertidal (Bousher, 1999; Griffiths et al., 2018; Griffiths et al., 2019a; Lloyd et al., 2013; Wooldridge et al., 2017b; Wooldridge et al., 2018). A central basin, protected by two barrier spits (Drigg Spit to the north and Eskmeals Spit to the south), is fed by three main rivers, the Rivers Irt, Mite, and Esk, which flow westwards into the Irish Sea. The Holocene valley fill succession sits upon Devensian glacial diamicton, that is directly overlain by fluvial gravels or peat beds (Coleman et al., 2021; McGhee et al., 2022; Merritt & Auton, 2000). The estuary formed in five stages over a 12,000 year period: (1) Late Devensian lowstand and valley incision (ca. 12,000 – 10,500 yrs. bp); (2) Early Holocene rapid transgression (ca.

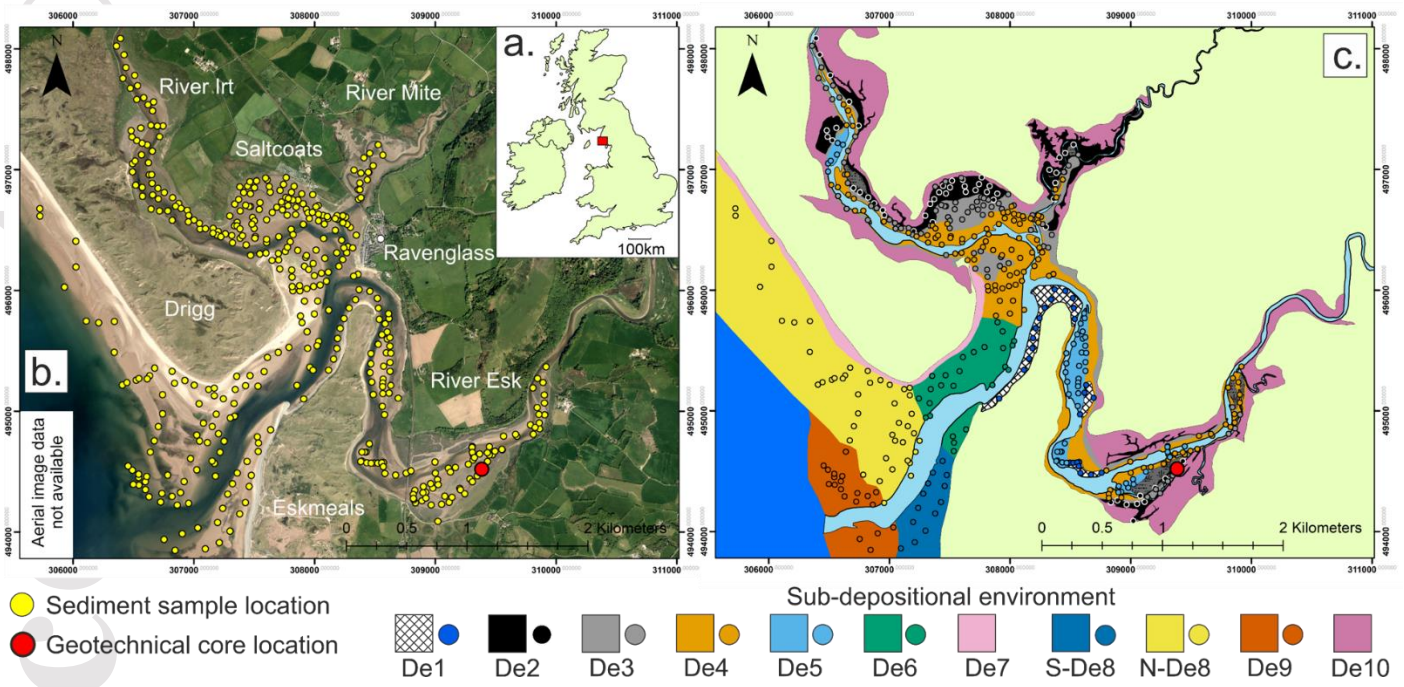

Figure 1. – a) location map with a red square indicating the location of the Ravenglass estuary, Cumbria, NW England; b) overview map of the Ravenglass Estuary. Surface sample locations (>2 cm depth) are indicated using yellow dots; c) a map of present day sub-depositional environments in the Ravenglass Estuary (modified after Simon et al., 2021). All surface samples are shown, colored by their corresponding sub-depositional environment.

10,500 – 6000 yrs. bp); (3) Holocene high-stand (ca. 6,000 – 5,000 yrs. bp); (4) minor fall in relative sea level (ca. 5000 yrs. bp to 410 yrs. bp); and, (5) extensive backfill of the Irt and Esk channels and merging into central basin (ca. 410 – present yrs. bp) (Lloyd et al., 2013; McGhee et al., 2022). Mineralogically, sediment in the estuary is dominated by quartz, with variable quantities and assemblages of feldspars, micas, ferromagnesian minerals and clay minerals, reflecting the varied hinterland geology (Daneshvar & Worden, 2018; Griffiths et al., 2019a; Griffiths et al., 2019b; Wooldridge et al., 2017b; Wooldridge et al., 2018; Wooldridge et al., 2019b).

The surface sediments of the Ravenglass Estuary have been extensively studied, with geomorphologically distinct sub-depositional environments identified using key diagnostic features, such as bedforms, sedimentary structures, sediment texture, and bioturbation traces (i.e., lithofacies and ichnofacies) (Daneshvar & Worden, 2018; Griffiths et al., 2018; Griffiths et al., 2019a; Griffiths et al., 2019b; Muhammed et al., 2022; Simon et al., 2021; Wooldridge et al., 2017a; Wooldridge et al., 2017b; Wooldridge et al., 2018; Wooldridge et al., 2019a). Modern sub-depositional environments in the Ravenglass Estuary have been defined (Figure 1b) and include gravel beds (De1), tidal flats (mud flats (De2), mixed flats (De3), sand flats (De4)), tidal bars (De5), tidal inlet (De6), backshore (De7), foreshore (De8), pro-ebb delta (De9), and saltmarsh (De10) (Griffiths et al., 2018; Simon et al., 2021; Wooldridge et al., 2017b). Tidal flat sediments are defined here using textural data based on a modified tidal flat classification scheme proposed by Brockamp and Zuther (2004) after Reineck and Siefert (1980), where mud flat (De2) contains 15 to 50% sand, mixed flat (De3) contains 50 to 90% sand, and sand flat (De3) contains >90% sand. The mapped distribution of sub-depositional environments in the Ravenglass Estuary is displayed in Figure 1b. The foreshore sub-depositional environment (De8) has been split into northern foreshore (N-De8) and southern foreshore (S-De8), as there is a significant textural and compositional difference between these sub-depositional environments owing to complex sediment movement patterns within the estuary (Muhammed et al., 2022; Simon et al., 2021).

Where sub-depositional environments are texturally similar, and laterally adjacent, they have been merged to form estuarine zones. Here, sand flat (De4) and tidal bar (De5) have been grouped to form an inner-estuary-coarse zone, and tidal inlet (De6), foreshore (N-De8 and S-De8), and pro-ebb delta (De9) have been grouped to form an outer-estuary zone. The aim of grouping sub-depositional environments is to improve the prediction accuracy of machine learning models, for example Muhammed et al. (2022) found that merging De4 and De5 improved model accuracy by 7.5 %. This grouping is comparable to traditional facies models frequently used in sedimentology (Boyd et al., 2006; Dalrymple et al., 1992).

3 Materials and Methods

3.1 Overall workflow

The supervised machine learning workflow developed here employs Extreme Gradient Boosting (XGBoost) and Bayesian Optimization (BO) to classify estuarine sub-depositional environment using sediment texture. Hyperparameters are numerical values, used by the XGBoost algorithm, that affect model training and therefore overall model accuracy. Hyperparameters require tuning in order to return the best performing model. BO is a smart and efficient method of hyperparameter tuning for machine learning algorithms that builds a probability model and uses the results to select and test the combination of hyperparameter values with highest chance of improvement to build on what BO algorithm already knows. BO is an iterative process, and the number of iterations is controlled by the user.

We have used a training dataset of surface sediment texture data to calibrate XGBoost models. We present two different models in order to predict a suite of eight surface sub-depositional environments (Model 1) and four surface estuarine zones (Model 2). The surface calibrated models have been applied to unseen subsurface core data from the Ravenglass Estuary to classify sub-depositional environment (Model 1) and estuarine zone (Model 2) in the Holocene succession.

3.2 Surface textural data: training dataset

For this study, we used the surface dataset described by Simon et al. (2021) to calibrate an XGBoost machine learning workflow. This full dataset comprises 482 surface sediment samples (where surface is defined as < 2 cm depth) collected throughout the Ravenglass Estuary (Figure 1b). Textural data from surface sediment samples were collected using a Beckman-Coulter LS13-320 laser particle size analyzer (LPSA), and processed using GRADISTAT version 9.1 to obtain textural attributes (Blott & Pye, 2001). Nineteen textural attributes were used as predictors in the XGBoost algorithm (mean grain size (µm), sorting (µm), skewness, kurtosis, and primary modal grain size (µm), and volumetric percentages of very

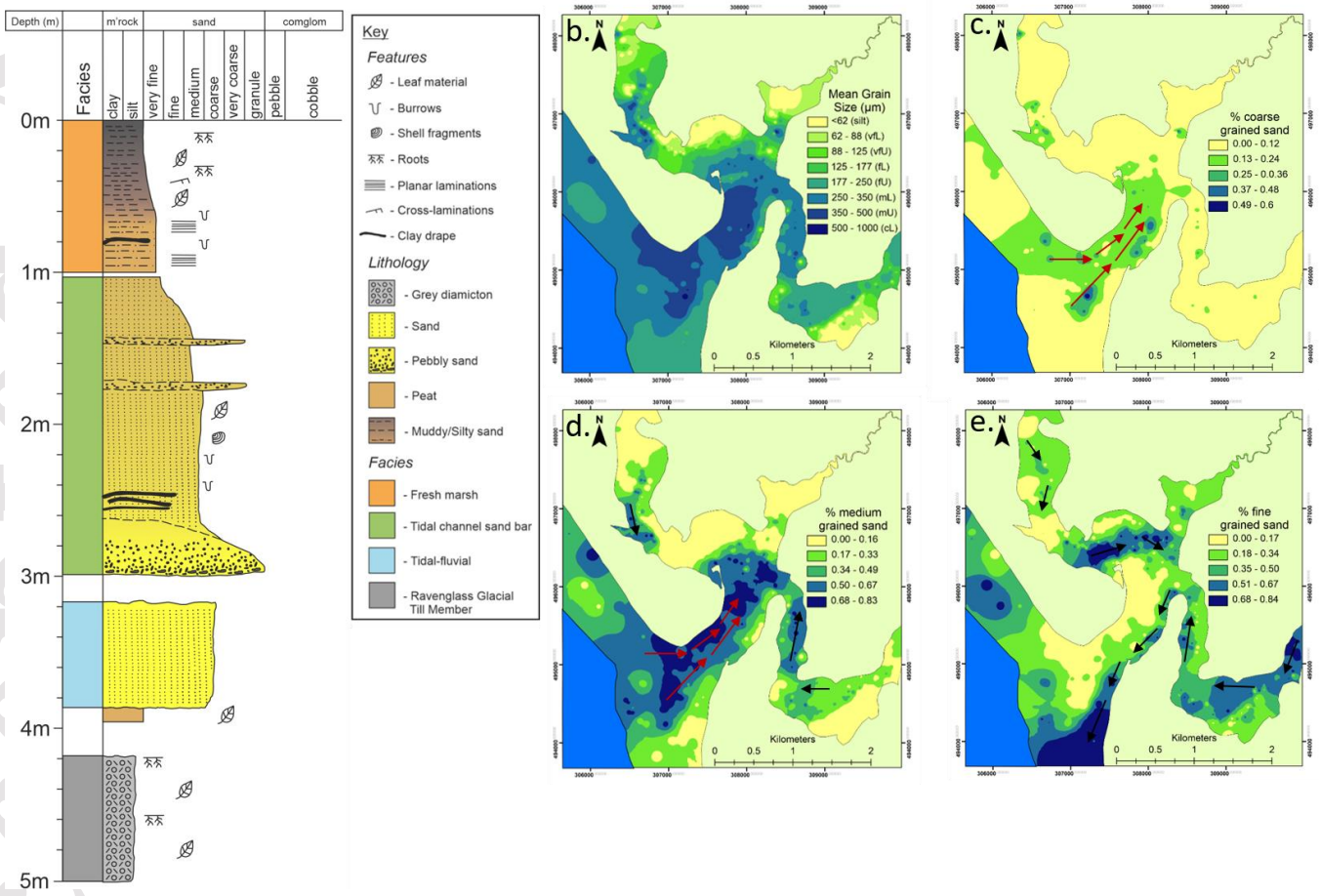

Figure 2. a) graphic sedimentary and facies log of the geotechnical core (Modified after McGhee et al., 2021) drilled in the Ravenglass estuary (location is indication figure 1b). b-f) grain size distribution plots in the Ravenglass Estuary. b) Mean grain size (µm). vfL = Lower Very Fine sand, vfU = Upper Very Fine sand, fL = Lower Fine sand, fU = Upper Fine Sand, mL = Lower Medium sand, mU = Upper Medium sand, cL = Lower Coarse sand. Arrows (black = fluvial & red = marine) indicate the direction of net sediment transport within the Ravenglass Estuary (modified after Simon et al., 2021).

coarse sand, coarse sand, medium sand, fine sand, very fine sand, very coarse silt, coarse silt, medium silt, fine silt, very fine silt, clay, silt, sand, and mud). Each surface sediment sample was assigned to a sub-depositional environment following the classification scheme proposed by Griffiths et al. (2018). For the purposes of this study, a subset of 435 samples from the full surface dataset was used as these belong to the eight sub-depositional environments, not all of which can be discriminated by visual inspection (gravel bed, De1, and saltmarsh, De10, samples were not included in the XGBoost model as these can easily be identified in core). Backshore (De7) sediment samples were not included due to the low sample size (n = 7) (Simon et al., 2021). In this study the new XGBoost workflow will be implemented to predict two factors, sub-depositional environment (Model 1), and estuarine zone (Model 2; see section 2.1), resulting in the production of two predictive models. 3.3 Holocene core: recovery and sampling

A geotechnical core (Figure 2a) was recovered from the Ravenglass Estuary in 2016, through the full thickness of the Holocene estuarine succession, from a vegetated tidal bar in the Esk River (Figure 1b). Core retrieval took place in one-meter segments using a Geotechnical 'P60' rotary rig and was stored in a semi-rigid plastic liner for transportation. To obtain textural data via LPSA, sediment samples (approximately 20 g) were extracted from the air-dried, cut-face of the core at five-centimeter intervals, ensuring that only one lithological type was sampled for each interval. If a sampling interval fell on a sharp lithological boundary (e.g., mud and sand) a sample was taken each side of the boundary.

McGhee et al. (2022) interpreted the geotechnical core using detailed traditional descriptive sedimentary logging, radiocarbon dating, and facies analysis. McGhee et al. (2022) interpreted the base of the core to belong to the Ravenglass Till Member (RGTM). The RGTM is gray to red, very fine grained, and very poorly sorted, with some pebbles and shell fragments. This is overlain, from 400 cm to 320 cm, by massive tidal-fluvial sands, which are fine to medium grained and moderately to well sorted. Sedimentary structures are uncommon, but silty laminae are recorded. The core from 300 cm to 100 cm was interpreted to be a package of tidal channel sand bar sediment. This package is composed of fine to medium sand, with pebbles at the base of the package. The interval is poorly sorted at the base, but moderately-well sorted upwards. Clay drapes, flaser bedding and silt-rich laminae are common. The

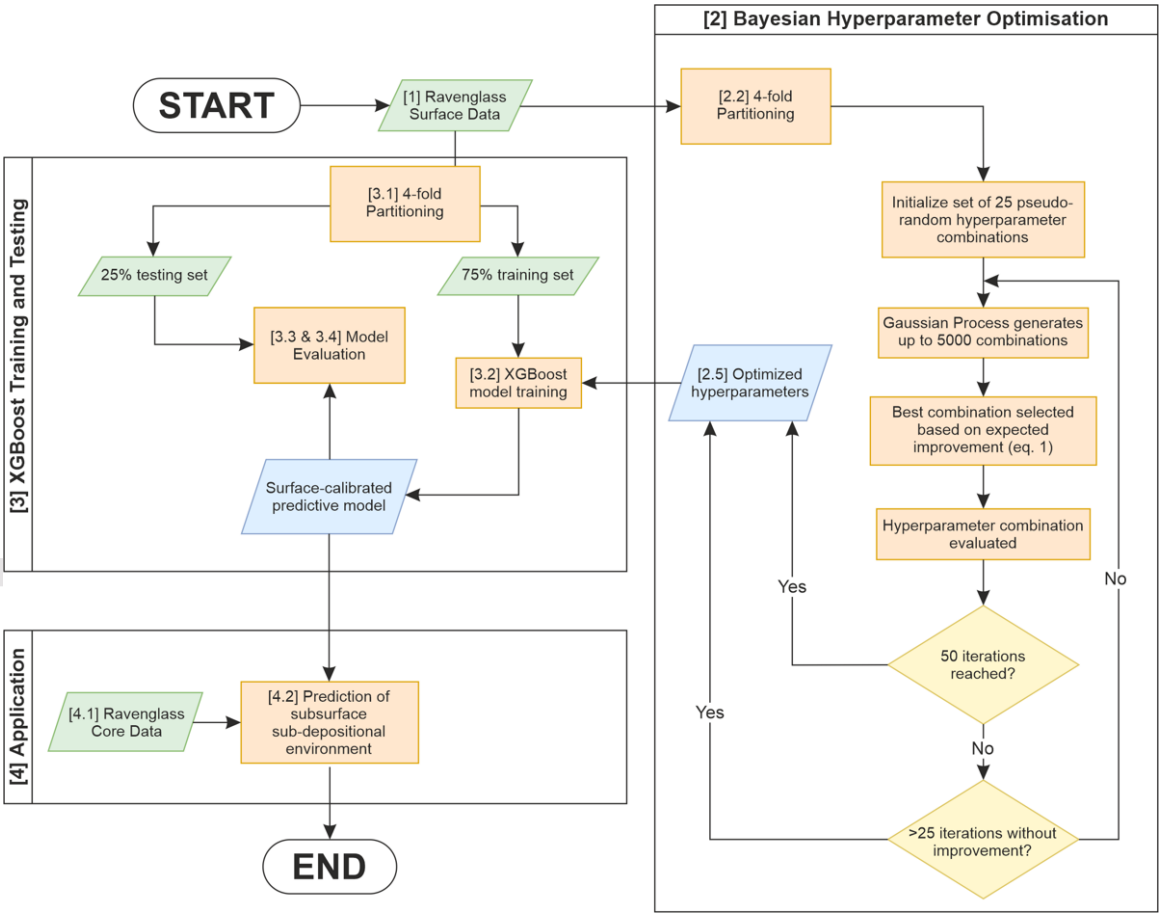

Figure 3. Machine learning workflow implemented in this study (adapted after Sun et al., 2020). Bayesian Hyperparameter Optimization must be completed before XGBoost training and testing. Green rhomboids indicate input data while blue rhomboids indicate data generated internally by the workflow. Numbers in square brackets indicate code section where the step occurs.

channel sand bar is capped by an interval of fresh-marsh from 100 cm to 0 cm. This surface package is composed of very fine sand and silt laminations and fines upwards. Roots are common in the top 20 cm.

3.4 Automated classification of sub-depositional environment

3.4.1 Overview of Machine Learning Workflow

The machine learning workflow (Figure 3) developed here has been written in RStudio (R Core Team, 2016) implementing the 'XGBoost' and 'Tidymodels' packages, and can be used with any numerical input data for classification problems (Chen & Guestrin, 2016; Wickham et al., 2019). All software and packages used in the workflow are open source and freely available. The XGBoost workflow code can be found in the repository (https://doi.org/10.5281/zenodo.7003918), which contains the necessary data and code to reproduce the results presented in this study.

The workflow is built from five sections of code (Figure 3). The sections consist of setting up the RStudio project (code section 0), importing surface sediment textural data (code section 1), optimizing XGBoost hyperparameters using Bayesian Optimization (code section 2), model training and testing (code section 3), and application to subsurface sediment textural data from the geotechnical core (section 4).

3.4.2 Bayesian Optimization of Hyperparameters: code section 2

Hyperparameters are numerical values within the XGBoost algorithm that control how a model is built, for example controlling what fraction of the training data is used to build a model, or the maximum number of data-splits a decision tree can have (Chen & Guestrin, 2016). The hyperparameters have default values; however, these may not be suitable for every data type and classification problem, leading to unrepresentatively poor model performance purely because of unsuitable (non-optimised) hyperparameter values.

Table 1. Descriptions, search ranges, and optimum (tuned) values of each hyperparameter tuned within the workflow (Figure. 2). Optimum values are shown for Model 1 and Model 2.

Hyperparameter name	Hyperparameter number	Description	Minimum	Maximum	Model 1 Optimum	Model 2 Optimum
Nrounds	1	Maximum number of boosting iterations.	10	1000	336	605
Max_depth	2	Maximum depth allowed for each tree. A higher value will lead to a more complex model.	4	16	16	16
Min_child_weight	3	Minimum number of samples in a node where the algorithm can try to partition further. If there are fewer than the minimum value at the node, then the node becomes a leaf.	1	20	1	1
Gamma	4	A node is split when the resulting split would give a positive reduction in the loss function. Gamma is the minimum loss reduction required to make further partition on a node. The larger the value of gamma, the more conservative the model will be.	0.01	3	2.971805	1.261678
Eta	5	Learning rate, where lower values allow a more detailed model to be constructed which is resistant to overfitting but computationally intensive while a high value creates a model quickly.	0.001	0.1	0.013150	0.002590
Colsample_bytree	6	Number of columns/features from the dataset that will be used to train a model.	2	19	17	9
Subsample	7	Proportion of training data sampled to build trees on each boosting iteration.	0.6	1	0.762498	0.848946

Bayesian optimization (BO) was implemented as the method of hyperparameter tuning as it is highly efficient and one of the most advanced techniques in selecting hyperparameters in machine learning and artificial intelligence (Ghahramani, 2015; Joy et al., 2016). It has been suggested that BO leads to an overall higher model accuracy than other state-of-the-art hyperparameter tuning algorithms and is less

prone to overfitting data, i.e., less likelihood of the model picking up on noise or random fluctuations in the training data (Snoek et al., 2012). To find the optimum values for the seven tuneable hyperparameters within Tidymodels (Table 1), we have used a BO hyperparameter tuning method, modified after Sun et al. (2020). Hyperparameter tuning carried out by Sun et al. (2020), was undertaken only on one training dataset, whereas we have implemented 4-fold partitioning (four independent testing and training datasets with a 25:75 split extracted from the whole dataset) of the data to produce hyperparameter combinations that are unbiased and generalized for the entire dataset.

The initial Ravenglass surface dataset was split into 4-folds (splits) for cross-validation to reveal whether selection bias or overfitting have occurred. An initial set of 25 pseudo-random hyperparameter combinations was generated using Latin hypercube sampling of pre-defined ranges of the seven hyperparameters and were evaluated for accuracy. Hyperparameter search ranges were initially based on those from Sun et al. (2020), but reduced after further testing to remove unfavourable values. Next, a Gaussian Process Model (which describes the distribution of model accuracy with respect to hyperparameters) used the initial set of 25 pseudo-random hyperparameter combination as predictors to generate up to 5000 further candidate hyperparameter combinations, as this is the maximum that the function allows (MacDonald et al., 2015). An acquisition function (Equation 1) (https://tune.tidymodels.org/articles/acquisition_functions.html) identified the candidate hyperparameter combination with the highest probability of improvement over the previous best combination and selected it for evaluation using the 4-fold cross-validation.

For a given hyperparameter combination, θ, there is a predicted mean metric, $\mu(\theta)$, and an associated error for that metric, $\sigma(\theta)$. The previous best mean performance is also known, m_{opt}. A trade off term (τ) is used to allow for exploration of hyperparameter combinations, rather than exploiting tight search ranges (Kuhn, 2022). The expected improvement for a given combination is determined using:

$$EI(\theta; m_{opt}) = \delta(\theta)\Phi\left(\frac{\delta(\theta)}{\sigma(\theta)}\right) + \sigma(\theta)\phi\left(\frac{\delta(\theta)}{\sigma(\theta)}\right) \quad \text{Equation 1}$$

Where:

$$\delta(\theta) = \mu(\theta) - m_{opt} - \tau;\ \tau = 0.01$$

$\Phi(\cdot)$ is the cumulative standard normal and $\phi(\cdot)$ is the standard normal density.

The BO process was iterated, updating the Gaussian Process Model between iterations, until either no improvement was made on overall accuracy after 25 iterations, or 50 total iterations were reached. These values have been increased over the default values (5 initial combinations, 10 iterations total) in order to obtain the most accurate model possible. These values can be increased according to the computational power and time available to the user; on our hardware (supplementary information Text S2) tuning could be completed within one hour. After this process, the hyperparameter combination with the best overall accuracy was determined to be the optimum combination and was extracted for subsequent application to the data (code section 2.5).

BO method of hyperparameter tuning is favoured over only a grid search (with potentially $>10^5$ combinations) or only Latin hypercube sampling that employs a user-defined number of combinations which evenly cover n-dimensional hyperparameter space. Random grid search is an alternative approach to hyperparameter tuning. This method selects and tests a user-defined number of hyperparameter values that are randomly distributed within the search space. The random grid search is less computationally-demanding than full grid search, but the optimum combination of hyperparameters might not be identified. BO uses previous attempts to identify which hyperparameter values produced the best accuracies for subsequent iterations to prefer, thereby not wasting computational resources on testing combinations that are unlikely to yield a high accuracy (Ghahramani, 2015).

3.4.3 Model Training and Testing: code section 3

To build and evaluate a final XGBoost model, the data were split 75:25 into 4-fold of training and testing data (i.e., four discrete splits in the data to train and then test the model four time). The training data and

optimum combination of hyperparameters, generated in code section 2, were used to build the final predictive model.

Table 2. Evaluation metrics utilized in this study with their defining equations and brief description. Each of these metrics' values can range from 0 to 1, with 1 being the 'best' score, and 0 the 'worst'. *TN* = 'True Negative', *FN* = 'False Negative', *TP* = 'True Positive', *FP* = 'False Positive'. Modified after Sokolova and Lapalme (2009).

Metric	Equation	Description
Specificity	$\frac{TN}{TN + FP}$	Measures the effectiveness of a classifier at identifying negative labels
Precision	$\frac{TP}{TP + FP}$	The proportion of positive prediction that are true positives
Recall	$\frac{TP}{TP + FN}$	Measures the effectiveness of a classifier at identifying positive labels
F1 Score	$\frac{TP}{TP + \frac{1}{2}(FP + FN)} = 2 \cdot \frac{\text{Precision} \times Recall}{\text{Precision} + Recall}$	A measure of overall accuracy which does not take into account true negatives
Balanced Accuracy	$\frac{Specificity + Recall}{2}$	The mean of specificity and recall, balances the rate of true positives and true negatives
Overall Accuracy	$\frac{\sum_{i=1}^{l} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}}{l}$	Overall effectiveness of a classifier, where 'l' is the number of classes

The final predictive models were applied to the testing dataset for evaluation. Mean and standard deviation were reported for each performance metric. The accuracy of the individual sub-depositional environments was used to compare performance between classes within a model, whereas overall accuracy was used to compare between models. To measure the performance of sub-depositional environments, we have used specificity, recall, precision, F1 score, and balanced accuracy (Table 2; Sokolova & Lapalme, 2009; Tharwat, 2021).

3.4.4 Application of the workflow to core data: code section 4

Calibrated XGBoost models, produced in code section 3.2, were applied to textural data from a geotechnical core drilled through a present-day tidal bar (Figure 2) in order to classify sub-depositional environment in the subsurface of the Ravenglass Estuary. The output of this section includes a stacked bar plot of the probability that an interval is a particular sub-depositional environment, and a log of the most probable sub-depositional environment.

4 Results

4.1 Model Tuning and Optimization

Bayesian Optimization (BO) uses accuracy data from the previous iteration during the tuning process, and therefore the optimum value for each hyperparameter is found with relatively few iterations (16 for Model 1 and 16 for Model 2). A visual example of the BO tuning process for Model 1 is shown in Figure S1 (supplementary information). Improvement can also be seen in the ranges and means of accuracy from initial and iteration combinations where the range of accuracy is both higher and narrower for iterations when compared to initial values when tuning both models.

The BO algorithm reached the optimum values for Model 1 at iteration 16 (out of 41 total iterations) and for Model 2 at iteration 16 (out of 41 total iterations). Optimized hyperparameter values for each model and their search ranges are displayed in Table 1. Hyperparameter numbers 4 and 5 were selected to be searched on a logarithmic scale as their ranges span several orders of magnitude. Model 1 initial hyperparameter combination accuracies range from 49.6% to 69.5% with a mean of 64.7%, whereas iteration accuracies range from 62.3% to 71.3% with a mean of 68.8%. Model 2 initial hyperparameter combination accuracies range from 75.4% to 83.4% with a mean of 80.2%, whereas iteration accuracies range from 77.7% to 84.6% with a mean of 82.9%.

Both Model 1 and Model 2 show the same values for hyperparameters 2 and 3. This suggests these hyperparameters have a controlling factor on how the predictive models perform on the dataset presented here. Values for all other hyperparameters vary between Model 1 and Model 2. This difference suggests that these hyperparameters do not play a dominant role in the overall accuracy of the predictive models for the dataset presented here. Rijn and Hutter (2018) conducted an extensive comparison of the importance of hyperparameters across datasets for different machine learning models that included Random Forest and ADA boost. Rijn and Hutter (2018) found that hyperparameter 3 ('min_samples_leaf' in Random Forest) and hyperparameter 6 ('max_features' in Random Forest) lead to the highest degree of variation in model accuracy. Here, our results support the finding that hyperparameter 3 leads to a large variation in model accuracy, however hyperparameter 6 does show the same trend.

BO increases the classification performance of both Model 1 and Model 2. Similar improvements in model performance when employing BO have been noted across a range of classification disciplines that include the prediction of advance rate of tunnel boring machine under hard rock conditions (Zhou et al., 2021), prediction of undrained shear strength in soft clays (Zhang et al., 2021), and the spatial prediction of shallow landslides (Kavzoglu & Teke, 2022).

4.2 Surface-calibrated XGBoost Model Evaluation

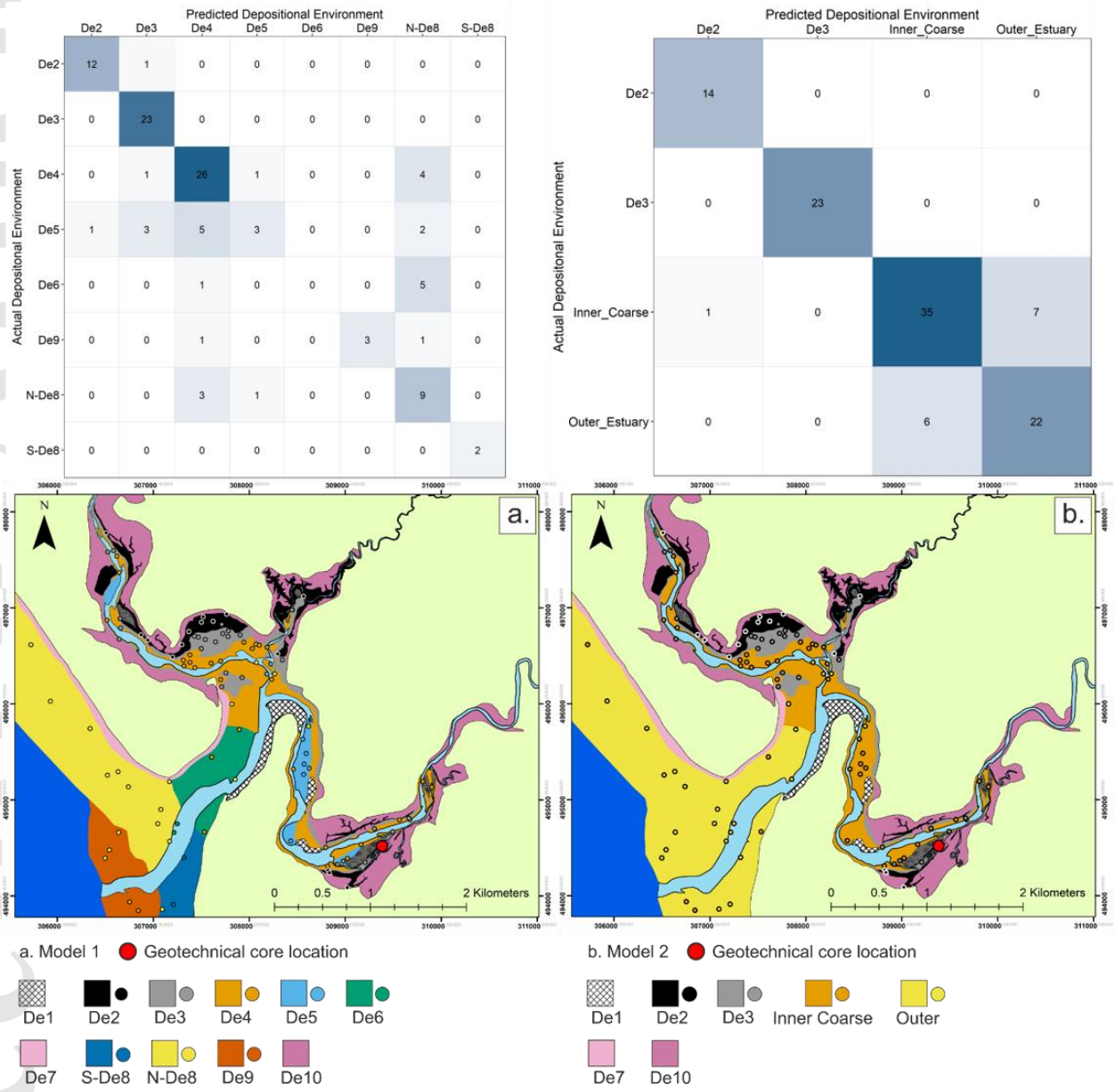

Figure 4. a) Confusion matrix for model 1 (sub-depositional environment). A confusion matrix is used to evaluate the performance of a predictive model. All data displayed in a confusion matrix is part of the testing data set, fold 3. The y-axis indicates the actual sub-depositional environment of a point, and the x-axis indicates the predicted sub-depositional environment of a point (e.g. considering De2, 14 points have been predicted as De2 and 1 point has been predicted as De3). Overall accuracy for the testing data shown is 72.48%. b) A confusion matrix for model 2 (estuarine zone). Model 2 has an overall accuracy for data shown is 85.32%. c) Model 1 testing data colored by predicted environment with mapped sub-depositional environments; d) Model 2 testing data colored by predicted environment with mapped merged sub-depositional environment.

Table 3. Evaluation metrics for Model 1 derived from testing data, averaged over 4-folds (1 standard deviation in parentheses). Overall Accuracy = 68.96% (1sd = 2.10%).

Metric	Equation	Description
Specificity	$\frac{TN}{TN + FP}$	Measures the effectiveness of a classifier at identifying negative labels
Precision	$\frac{TP}{TP + FP}$	The proportion of positive prediction that are true positives
Recall	$\frac{TP}{TP + FN}$	Measures the effectiveness of a classifier at identifying positive labels
F1 Score	$\frac{TP}{TP + \frac{1}{2}(FP + FN)}$ $= 2 \cdot \frac{\text{Precision} \times Recall}{\text{Precision} + Recall}$	A measure of overall accuracy which does not take into account true negatives
Balanced Accuracy	$\frac{Specificity + Recall}{2}$	The mean of specificity and recall, balances the rate of true positives and true negatives
Overall Accuracy	$\frac{\sum_{i=1}^{l} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}}{l}$	Overall effectiveness of a classifier, where 'l' is the number of classes

Evaluation metrics for Model 1 and Model 2 are presented in Tables 3 and 4 respectively. These metrics have been averaged for the 4-folds. Figure 4 shows data for one test/train permutation (fold 3) as data for multiple permutations cannot be shown on these graphs. Classification data are presented for Model 1 (Figure 4a) and Model 2 (Figure 4b) as confusion matrices. These display the predicted class compared to the actual class for all testing data points (i.e., the remainder of the data not used in the model calibration: n = 109) and allow visualization of where correct and incorrect classifications have occurred. The testing data shown in Figures 4a and 4b are presented spatially in Figures 4c and 4d. Here, the testing data points have been colored by the predicted sub-depositional environment and plotted on a map of surface sub-depositional environments in the Ravenglass Estuary (Figure 1c).

Table 4. Evaluation metrics for Model 2 derived from testing data, averaged over 4-folds (1 standard deviation in parentheses). Overall Accuracy = 84.14% (1sd = 1.92%).

Metric	Equation	Description
Specificity	$\frac{TN}{TN + FP}$	Measures the effectiveness of a classifier at identifying negative labels
Precision	$\frac{TP}{TP + FP}$	The proportion of positive prediction that are true positives
Recall	$\frac{TP}{TP + FN}$	Measures the effectiveness of a classifier at identifying positive labels
F1 Score	$\frac{TP}{TP + \frac{1}{2}(FP + FN)}$ $= 2 \cdot \frac{\text{Precision} \times Recall}{\text{Precision} + Recall}$	A measure of overall accuracy which does not take into account true negatives
Balanced Accuracy	$\frac{Specificity + Recall}{2}$	The mean of specificity and recall, balances the rate of true positives and true negatives
Overall Accuracy	$\frac{\sum_{i=1}^{l} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}}{l}$	Overall effectiveness of a classifier, where 'l' is the number of classes

4.3 Predicting paleo-sub-depositional environment in geotechnical core

The two models calibrated using surface sediment texture have been applied to textural data from the Holocene core in order to predict sub-depositional environment in the subsurface (Figure 5). Presented here, alongside the graphic sedimentary log (Figure 5a), are two probability bar charts (Figure 5b & c) which capture the uncertainty of sub-depositional environment prediction by the XGBoost model in the subsurface. The predicted environment plot accompanying this represents the most probable sub-environment for a sample interval. The unbiased predictions of an XGBoost model enhance the interpretations that can be made by traditional sedimentary logging.

From 500 cm to 420 cm in the geotechnical core, grey diamicton has been visually discriminated as basal till (RGTM) due to its very poorly sorted, clay-rich nature. Model 1 predictions (Figure 5b) show mixed sand flat (De4) with tidal bar (De5) from 400cm up to 320 cm depth. Pebbly sand between 290 – 275 cm can be visually discriminated by the presence of gravel and pebbles and so should be considered to be gravel bed (De1). Samples between 270 – 120 cm depth are mainly predicted as sand flat (De4) with two points predicted as N-foreshore (N-De8) and mixed flat (De3). From 95 cm to the shallowest sample at 20 cm, samples are predicted as mixed flat (De3) and mud flat (De2) with mixed flat (De3) predictions dominant towards the bottom of this sediment package. Due to abundant root material in the top 20 cm of the core, these intervals were not sampled for textural analysis and discrimination of sedimentary environment as they could be visually identified as saltmarsh (De10). A comparison of the predicted environment logs between Model 1 and Model 2 reveals a significant correlation between prediction made by Model 1 corresponding to the zone predicted by Model 2. The predictions made by the models are able to resolve subtle details, particularly in the sand-dominated sub-depositional environments, that had not been identified during traditional descriptive sedimentary logging and facies interpretation.

5 Discussion

In the following sections we evaluate the machine learning workflow presented here with respect to the effectiveness and reliability of model predictions of subsurface sub-environments from 'unseen' textural information of sedimentary core. An outline of the key limitations and uncertainties of the XGBoost machine learning approach is presented, as applied to the field of sedimentology. This will provide a basis for discussion on how the XGBoost workflow can be applied to other classification problems.

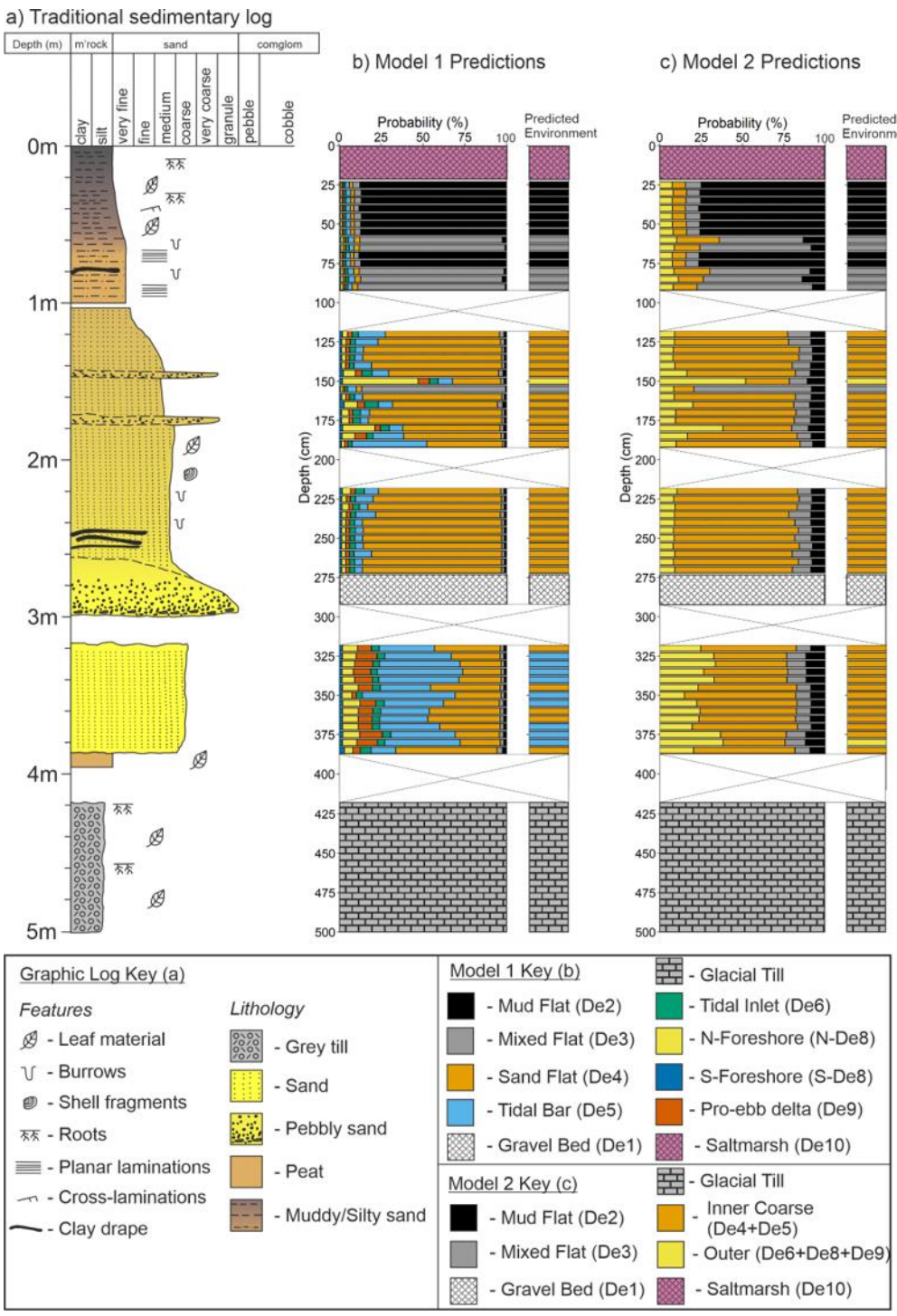

Figure 5. a) graphic log of a borehole through the Ravenglass Estuary succession (location indicated on Figure 1b, after McGhee et al,. 2021). XGBoost prediction probability logs and environment with highest probability ('Predicted Environment') for: b) Model 1 and; c) Model 2.

5.1 Evaluation and Interpretation of Predictive Models

Presented are two models for the automated classification of estuarine sub-depositional environment using textural data. Model 1 has an average overall accuracy of 68.96% and is presented to assess the viability for the discrimination of all sub-depositional environments without mergers, and presents an opportunity to discuss the differences and similarities between sub-depositional environments. In contrast, Model 2 is able to discriminate estuarine zones and has a higher average overall accuracy of 84.14%. We suggest how the models should be interpreted when applied to unseen data, such as that from a geotechnical core drilled in the same environment from which the surface samples were collected. The evaluation metrics of Model 1 and Model 2 are presented in Table 3 and Table 4.

In both models, mud flat (De2; 15 to 50% sand) and mixed flat (De3; 50 to 90% sand) were pre-defined using textural data using a modified tidal flat classification scheme proposed by Brockamp and Zuther (2004) after Reineck and Siefert (1980). Due to the mathematical definition of these sub-depositional environments, it is not surprising that both are predicted with a high accuracy in both Model 1 and Model 2. The drive behind the modelling we present however was not to distinguish the sediments from De2 and De3 but instead those of the sand-rich environments (De4-De9) which dominate the estuary at the surface and in the cored sediment from the subsurface.

5.1.1 Evaluation of Model 1: Classifying sub-depositional environments

Sand flat (De4) sediment is defined using the tidal flat classification scheme from Brockamp and Zuther (2004) based on more than 90% sand content. All sand-dominated sub-depositional environments in the Ravenglass Estuary share this characteristic, but are differentiated at the surface by sedimentary structures and ichnofabrics. De4 has a precision score of 0.597, indicating that approximately 40% of samples predicted as De4 are incorrect. Examination of the confusion matrix for Model 1 (Figure 4a) shows that samples from all sand-dominated environments (i.e., not De2 or De3) may be predicted as De4. Mispredictions may be because the sub-depositional environments are adjacent and therefore there is a high likelihood that sediment mixing occurred between these environments. De4 has a recall score of

0.757, showing that only 25% of known De4 samples are incorrectly predicted as other sub-depositional environments. Some De4 samples are incorrectly predicted as N-De8. These sub-depositional environments are not laterally adjacent, this therefore suggests that sediment movement between these two sub-depositional environments may be facilitated in De6 as both De4 and N-De8 are adjacent to De6. Interpreted sediment movement patterns (Figure 2c - e) demonstrate that the De6 sub-depositional environment acts as pathway between outer estuary and inner estuary environments, facilitating the movement of medium and coarse-grained sand into the central basin from the foreshore. The movement of sediment between the outer estuary and central basin may explain the mispredictions between De4 and N-De8. The confusion between De4 and N-De8 results in false positives in N-De8 predictions, causing it to have low specificity and precision scores (0.905, 0.494 respectively).

Tidal bar (De5) sediment has the second lowest Balanced Accuracy score at 0.588. This score is mainly due to the low recall score for this class (0.173) which contrasts with the high specificity score (0.958). The recall score indicates that less than a quarter of the actual De5 samples are correctly predicted in the surface sediment calibration dataset. Other samples are variably predicted as De4 and N-De8. This points towards textural similarity between these environments. The specificity score for De5 is high (0.958) with two false positives shown in the confusion matrix (Figure 4a). However, poor average recall, precision, and F1 scores suggest this environment cannot be reliably classified. The poor performance of an XGBoost model to predict the De5 sub-depositional environment may be due to a number of reasons; De5 is incorrectly defined in some places at the surface, De5 is highly variable and therefore has no consistent textural signature to define it, and De5 may have different textural signatures in the River Esk and River Irt where we have chosen group all tidal bars into one sub-depositional environment.

Tidal inlet (De6) sediment is never correctly predicted (Figure 4a). However, a high specificity value (0.995) shows that other environments are rarely predicted as De6. Simon et al. (2021) suggest that there is not a consistent textural basis upon which to identify sediment from the De6 sub-depositional environment. Samples that have been labelled as De6 are predicted as either De4 or N-De8 (both laterally

adjacent environments), further indicating the possibility that De6 is a transition or mixing zone between inner and outer estuary sediment (Figure 2c - e). Grain size increase basinward and decrease landward is driven by strong tidal flood currents that promote sediment transport, and lead to a mixing zone within the tidal inlet (McGhee et al., 2022). This interpretation is supported by the work of Simon et al. (2021) who identified an influx of medium to coarse sand along the northern tidal inlet and an efflux of fine sand to silt in the southern tidal inlet (Figure 2c - e).

Test data from Model 1 (fold 3) are represented spatially in Figure 4c, which shows that, within the upper arms of all rivers, there are no points predicted as outer estuary environments. However, moving toward the central basin, mispredictions are more common, with four incorrectly predicted points (De4 predicted as De5 and N-De8) at the confluence of the River Irt and River Mite. Incorrect predictions at the confluence may be due to this area representing a zone of bedload convergence and therefore deposition of more coarse sediment (Dalrymple et al., 1992). Figures 2c and 2d show that this confluence may be the limit of sediment influx into the central basin. Tidal forces directly interferes with sediment transport with an estuary by introducing sediment from external sources (e.g., transport marine sediment into an estuary during a flood tide) (Dalrymple et al., 1992). In estuarine systems with weaker tidal influence this affect would not be as strong and so sub-depositional environments may be more predictable.

5.1.2 Evaluation of Model 2: Classifying estuarine zone

The merging of sub-depositional environments improves the overall accuracy of the model, but there is a trade-off between accuracy and the ability to resolve sub-depositional environments. When merging sub-depositional environments, the resolution of the final model and application of the model to the interpretation of data from core will be reduced. This was implemented by Muhammed et al. (2022) where sub-depositional environments were merged to improve the accuracy of a predictive RPART model using geochemical data in the Ravenglass Estuary. The user of this type of approach should

therefore consider the purpose of the data analysis and the impact that merging classification groups would have when interpreting the results of models.

Model 2 has an average overall accuracy of 84.14% and offers an overview of estuarine zones as it is able to differentiate coarse inner-estuary sediment from outer-estuary sediment. This model's predictions should be considered in conjunction with the results of Model 1. Evaluation metrics for Model 2 are presented in Table 4.

For Model 2, De2 and De3 have similar specificity, precision, and recall score to Model 1 (Tables 3 & 4). This suggests that merging environments to form inner-coarse and outer-estuary does not affect the predictive capability of the model for De2 and De3 sub-depositional environments. This outcome is likely because De2 and De3 were defined texturally following a tidal flat classification scheme (Brockamp & Zuther, 2004).

Figure 4b shows that Model 2 led to some mispredictions between inner-coarse and outer-estuary zones, reflecting the mispredictions found in Model 1 (Figure 4a). Nonetheless, both inner-coarse and outer-estuary zones have high F1 scores (0.806 and 0.786 respectively; Table 2 & 4) indicating that they can be reliably distinguished by Model 2.

Spatial representation of testing data for Model 2 (fold 3) are shown in Figure 4d, which reveals the quality of the prediction of estuarine zone within the Ravenglass Estuary. Within the arms of all three rivers and the central basin predictions of sub-depositional environment are consistent with Model 1 (Figure 4c). However, there are several data points that are incorrectly predicted in both models (one in sandflat adjacent to the tidal inlet, three at the confluence of the River Irt and Mite, and one at the northern end of a tidal bar in the River Esk). These data are consistently predicted as outer estuary sub-depositional environments in both models. Incorrect predictions in the sand flat immediately adjacent to the tidal inlet seem to be affected by the influx of coarse sand from the marine part of the sedimentary system (Figure 2c) possibly providing an explanation for the misprediction as N-De8 in Model 1.

5.2 Implications and considerations when applying a surface-calibrated model to core data

Core analysis and interpretation is critical for many industries (e.g., engineering and reservoir quality analysis) as it provides a direct and quantitative measurements of the subsurface (McPhee et al., 2015). Quaternary cores are typically unconsolidated and key diagnostic features are obscured, making interpretation challenging. The use of a surface-calibrated classification model allows unbiased interpretation of core data, and the ability to extract hidden and subtle details that might otherwise have been missed. The workflow presented here is designed to be used alongside, and not replace, conventional manual core analysis and interpretation.

The present-day environment at the surface where the geotechnical core was drilled is a vegetated tidal bar, adjacent to mud flats on the south side and the tidal channel on the north side. At the top of the geotechnical core, mud flat (De2) is predicted, underlying saltmarsh (De10) reflecting the interpretation made by McGhee et al. (2022). Given the very high evaluation metric scores (Tables 3 and 4) for these sub-depositional environments, there can be a high confidence that these intervals are predicted correctly.

Sand flat (De4) is predicted by Model 1 between 120-270 cm depths, which has a precision score of 0.719 suggesting there is a high probability this prediction is correct. Model 2 predicts that only inner-coarse intervals are present, suggesting that despite the shifting coastline (McGhee et al., 2022), this location has remained within the inner estuary since post-glacial deposition began, approximately 10,000 years ago (Figure 5c). One point of N-De8/outer-estuary is predicted however this coincides with a coarse-grained horizon. Foreshore sediments are typically more coarse than inner estuary tidal flats so this may explain the prediction shown here (Simon et al., 2021). This, combined with it being a single interval suggests it should not be interpreted as a short-lived incursion of foreshore but, instead, may represent a rare storm event that introduced coarse foreshore sediment into the inner estuary.

Some tidal bar (De5) intervals are predicted in the core between 320 and 390 cm, and, given the precision score of 0.6 (for the model presented in Figure 4a), it is likely that this sediment is texturally similar to

tidal bars currently present in the Ravenglass Estuary. It is therefore possible that this interval represents an early tidal bar present in the estuary soon after the Holocene transgression (McGhee et al., 2022). This predictive output is contrary to the interpretation of McGhee et al. (2022) who, based on classical descriptive core logging, interpreted the entire 100 to 300 cm interval in this core as representing channel initiation and formation of a 'tidal channel sand bar' and the underlying sand as 'tidal-fluvial'. One limitation of the modelling presented is that present-day thalweg sediments have not been sampled and analyzed and therefore cannot be predicted. Because of this, if thalweg deposits are preserved in a core, as McGhee et al. (2022) suggests, then it would not be identified by the models. It is likely that the gravel bed (De1) interval present in the core represents thalweg deposition which fines upwards into marginal sand flats.

Muhammed et al. (2022) used geochemical data, obtained by portable X-ray Fluorescence (XRF) analysis, as predictors for a simple Recursive Partitioning and Regression Trees (RPART) machine learning approach to predict sub-depositional environment in the Ravenglass Estuary. The XRF calibrated scheme was applied to the same geotechnical core described in this study. Between 320 and 390 cm the XRF-based model predicted sand flat (De4) sub-depositional environment with a high probability whereas Model 1 predicts tidal bar (De5) with a high probability. This suggests that, although the sediment is texturally similar to surface tidal bars (De5), it is also geochemically similar to surface sand flat (De4) which may mean sediments between 320 and 390 cm are not represented at the present day surface. At the base of this package of sand at ~370 cm depth, Muhammed et al. (2022) predicted the presence of outer estuary ebb-tidal delta (De9) sediment whereas Model 1 predicts the occurrence of inner estuary sand flat (De4) sediment, and Model 2 predicts inner coarse estuary sediment. The conflicting predictions at the base of the core may suggest that the interval is not consistently represented by a surface (modern) sub-deposition environment. Alternatively, the conflicting predictions may suggest the simple RPART model may not be reliable and that use of the geochemical classification requires careful attention. A dataset combining sediment texture and geochemical data obtained using XRF would likely

improve the predictive accuracy of XGBoost models due to a strong correlation between sediment grain size and mineralogy (Griffiths et al., 2019a).

The sediment between 400 and 500 cm in the core (Figure 5) was interpreted as glacial till due to its very fine, massive nature and minor pebble content (McGhee et al., 2022; Muhammed et al., 2022). The XGBoost models have not been applied to this interval because glacial till is not estuarine sediment and so has not been included in the training dataset. Therefore, the application of estuarine models to glacial till would be inappropriate. Similarly, the model has not been applied to peat at 390 to 400 cm.

Understanding the context of sediment in core is important and should be considered when applying predictive models. Predictive models should not be used in isolation, and instead should be used in conjunction with traditional sedimentary logging to extract hidden and subtle details of the varying sequence of sub-depositional environments from core that otherwise would have been missed.

We anticipate the proposed XGBoost workflow could be implemented on datasets from other marginal marine depositional systems to enhance the interpretation of their subsurface deposits. Ultimately, detailed interpretations of ancient, buried deposits could be made using models derived from analogous modern systems.

5.3 Evaluating the Application of the XGBoost Workflow

The accessibility of the software and code means the workflow is flexible and can be adapted to use any numerical or categorical data for classification. Here the workflow is used for the classification of sub-depositional environment in the Ravenglass Estuary. However, sub-depositional environment could be classified using textural data from the surface of other estuaries (e.g. the Gironde Estuary; Virolle et al., 2019) in order to develop a database of XGBoost models which could be compared to explore if and why certain estuaries build more effective models. Additionally, the workflow could be used to create a model incorporating calibration data from multiple estuaries. Datasets from other sedimentary environments

could also be used for classification. An input dataset could originate from any depositional system and could include geochemical data obtained by XRF, micropaleontology data, and mineralogical data, thus expanding the use of the workflow beyond estuarine systems. The use of input datasets, such as wireline data, could allow for lithology or facies prediction from well logs (e.g. Martin et al., 2021; Sun et al., 2020; Xie et al., 2018). Application to wireline would require the manual classification of core internals to be used as a calibration dataset, to create a model, which can subsequently be applied to an entire well. This is important for the correlation of vertically and laterally adjacent rock units and could prove highly efficient for large scale reservoir quality prediction where facies could be identified from wireline data to direct exploration in the fields of carbon capture and storage, nuclear waste management, and hydrocarbon exploration. The implementation of XGBoost allows the workflow to be highly scalable and applicable to large datasets with many input variables, and has been proven to build models up to ten times faster than other existing algorithms due to parallel and distributed out-of-core computation (Chen & Guestrin, 2016).

When applying the workflow to other datasets the user should make considerations about how it is implemented. Firstly, the search range for hyperparameter tuning should be considered (code section 2.3, lines 81 to 87). For example, here we limit the search range of hyperparameter number 6 between 2 and 19, as there are 19 input variables. For a dataset with 30 variables, the user might choose to adjust this search range, up to a maximum of 30. Secondly, the number of tuning iterations can be adjusted based on the user's experience; if Bayesian Optimization does not produce significantly higher accuracies than initial results then the number of tuning iterations or early stopping condition can be increased (code section 2.3, line 90 & 92).

6 Conclusions

Here we propose an efficient and highly adaptable machine learning workflow, using Extreme Gradient Boosting and Bayesian Optimization, for use in complex classification problems. We have implemented

this workflow in the classification of sub-depositional environments using surface textural data in the Ravenglass Estuary, Cumbria, northwest England.

1. Model 1 predicts eight sub-depositional environments (68.96 % overall accuracy) and Model 2 predicts four estuarine zones (84.14 % overall accuracy) at the surface of the Ravenglass Estuary. Model 1 highlights the difficulty in discriminating sand-dominated environments, however Model 2 can reliably discriminate inner-coarse sediment from outer-estuary sediment, demonstrating how both models should be used in conjunction.

2. Bayesian Optimization has been successfully implemented into the workflow and is effective at providing better hyperparameter combinations compared to Latin hypercube sampling, yielding improved mean accuracy combinations with a narrower range of accuracies (65.2% vs 69.4% mean accuracy for Model 1).

3. Influx of coarse sand through the northern side of the tidal inlet into the inner estuary may explain some incorrect classifications of inner estuary points. The sediment movement pattern shows how the tidal inlet (De6) acts as a transitional zone between inner and outer estuary and therefore seems to lacks a unique textural signature, resulting in poor model performance.

4. Using surface-calibrated models, we can make unbiased predictions of sub-depositional environment in Holocene cores from the Ravenglass estuary using textural data, aiding the interpretation of the subsurface. The presented models can identify subtle characteristics of sediment packages that could not be interpreted using traditional core logging and facies analysis techniques due to the massive nature of the sediments.

Acknowledgments

This work was undertaken as part of the Chlorite Consortium at the University of Liverpool (UK), sponsored by BP, Equinor, and DNO.

Open Research

All data and code to reproduce the work in the paper are available in the repository https://doi.org/10.5281/zenodo.7003918 (Houghton et al., in prep.).

References

Allen, J. R. L., & Thornley, D. M. (2004). Laser granulometry of Holocene estuarine silts: effects of hydrogen peroxide treatment. *The Holocene, 14*(2), 290-295. https://doi.org/10.1191/0959683604hl681rr

Biederman, E. W. (1962). Distinction of shoreline environments in New Jersey. *Journal of Sedimentary Research, 32*(2), 181-200.

Blott, S. J., & Pye, K. (2001). GRADISTAT: a grain size distribution and statistics package for the analysis of unconsolidated sediments. *Earth Surface Processes and Landforms, 26*(11), 1237-1248. https://doi.org/10.1002/esp.261

Bousher, A. (1999). *Ravenglass Estuary: Basic characteristics and evaluation of restoration options* (Restrad-Td, Issue.

Boyd, R., Dalrymple, R. W., Zaitlin, B. A., Posamentier, H. W., & Walker, R. G. (2006). Estuarine and Incised-Valley Facies Models. In *Facies Models Revisited* (Vol. 84, pp. 0). SEPM Society for Sedimentary Geology. https://doi.org/10.2110/pec.06.84.0171

Brockamp, O., & Zuther, M. (2004). Changes in clay mineral content of tidal flat sediments resulting from dike construction along the Lower Saxony coast of the North Sea, Germany. *Sedimentology, 51*(3), 591-600.

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA. https://doi.org/10.1145/2939672.2939785

Coleman, C. G., Grimoldi, E., Woollard, H., Holton, D., & Shevelan, J. (2021). Developing 3D geological and hydrogeological models for the Low Level Waste Repository site, west Cumbria, UK. *Quarterly Journal of Engineering Geology and Hydrogeology, 54*(2), qjegh2020-2026. https://doi.org/10.1144/qjegh2020-026

Dalrymple, R. W., Zaitlin, B. A., & Boyd, R. (1992, Nov). Estuarine facies models - conceptual models and stratigraphic implications. *Journal of Sedimentary Petrology, 62*(6), 1130-1146. <Go to ISI>://WOS:A1992JX76100016

Daneshvar, E., & Worden, R. H. (2018). Feldspar alteration and Fe minerals: origin, distribution and implications for sandstone reservoir quality in estuarine sediments. In P. J. Armitage, A. R. Butcher, J. M. Churchill, A. E. Csoma, C. Hollis, R. H. Lander, J. E. Omma, & R. H. Worden (Eds.), *Reservoir Quality of Clastic and Carbonate Rocks: Analysis, Modelling and Prediction* (Vol. 435, pp. 123-139). https://doi.org/10.1144/sp435.17

Flood, R. P., Orford, J. D., McKinley, J. M., & Roberson, S. (2015). Effective grain size distribution analysis for interpretation of tidal–deltaic facies: West Bengal Sundarbans. *Sedimentary Geology, 318*, 58-74.

Friedman, G. M. (1961). Distinction between dune, beach, and river sands from their textural characteristics. *Journal of Sedimentary Research, 31*(4), 514-529. https://doi.org/10.1306/74d70bcd-2b21-11d7-8648000102c1865d

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics, 29*(5), 1189-1232. http://www.jstor.org/stable/2699986

Friedman, J. H. (2002, 2002/02/28/). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*(4), 367-378. https://doi.org/https://doi.org/10.1016/S0167-9473(01)00065-2

Ghahramani, Z. (2015, 2015/05/01). Probabilistic machine learning and artificial intelligence. *Nature, 521*(7553), 452-459. https://doi.org/10.1038/nature14541

Griffiths, J., Worden, R. H., Wooldridge, L. J., Utley, J. E. P., & Duller, R. A. (2018). Detrital clay coats, clay minerals, and pyrite: a modern shallow-core analogue for ancient and deeply buried estuarine sandstones. *Journal of Sedimentary Research, 88*(10), 1205-1237.

Griffiths, J., Worden, R. H., Wooldridge, L. J., Utley, J. E. P., & Duller, R. A. (2019a, Apr). Compositional variation in modern estuarine sands: Predicting major controls on sandstone reservoir quality. *American Association of Petroleum Geologists Bulletin, 103*(4), 797-833. https://doi.org/10.1306/09181818025

Griffiths, J., Worden, R. H., Wooldridge, L. J., Utley, J. E. P., Duller, R. A., & Edge, R. L. (2019b). Estuarine clay mineral distribution: Modern analogue for ancient sandstone reservoir quality prediction. *Sedimentology, 66*(6), 2011-2047.

Heap, A. D., Bryce, S., & Ryan, D. A. (2004, 2004/06/01/). Facies evolution of Holocene estuaries and deltas: a large-sample statistical study from Australia. *Sedimentary Geology, 168*(1), 1-17. https://doi.org/https://doi.org/10.1016/j.sedgeo.2004.01.016

Joy, T. T., Rana, S., Gupta, S., & Venkatesh, S. (2016, 4-8 Dec. 2016). Hyperparameter tuning for big data using Bayesian optimisation. 2016 23rd International Conference on Pattern Recognition (ICPR),

Kavzoglu, T., & Teke, A. (2022, 2022/04/22). Advanced hyperparameter optimization for improved spatial prediction of shallow landslides using extreme gradient boosting (XGBoost). *Bulletin of Engineering Geology and the Environment, 81*(5), 201. https://doi.org/10.1007/s10064-022-02708-w

Kuhn, M. (2022). *tune: Tidy Tuning Tools*.

Lloyd, J. M., Zong, Y., Fish, P., & Innes, J. B. (2013). Holocene and Late-glacial relative sea-level change in north-west England: implications for glacial isostatic adjustment models. *Journal of Quaternary Science, 28*(1), 59-70.

MacDonald, B., Ranjan, P., & Chipman, H. (2015). GPfit: An R package for fitting a Gaussian process model to deterministic simulator outputs. *Journal of Statistical Software, 64*, 1-23.

Martin, T., Meyer, R., & Jobe, Z. (2021, 2021-June-24). Centimeter-Scale Lithology and Facies Prediction in Cored Wells Using Machine Learning [Methods]. *Frontiers in Earth Science, 9*. https://doi.org/10.3389/feart.2021.659611

Mason, C. C., & Folk, R. L. (1958). Differentiation of beach, dune, and aeolian flat environments by size analysis, Mustang Island, Texas. *Journal of Sedimentary Research, 28*(2), 211-226. https://doi.org/10.1306/74d707b3-2b21-11d7-8648000102c1865d

McGhee, C. A., Muhammed, D. D., Simon, N., Acikalin, S., Utley, J. E. P., Griffiths, J., Wooldridge, L. M., Verhagen, I. T. E., van der Land, C., & Worden, R. H. (2022). Stratigraphy and sedimentary evolution of a modern macro-tidal incised valley– an analogue for reservoir facies and architecture. *Sedimentology, 69*, 696-723.

McPhee, C., Reed, J., & Zubizarreta, I. (2015). Chapter 1 - Best Practice in Coring and Core Analysis. In C. McPhee, J. Reed, & I. Zubizarreta (Eds.), *Developments in Petroleum Science* (Vol. 64, pp. 1-15). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-444-63533-4.00001-9

Merritt, J. W., & Auton, C. A. (2000, Nov). An outline of the lithostratigraphy and depositional history of Quaternary deposits in the Sellafield district, west Cumbria. *Proceedings of the Yorkshire Geological Society, 53*, 129-154. <Go to ISI>://WOS:000165681800005

Mohamed, I. M., Mohamed, S., Mazher, I., & Chester, P. (2019). Formation Lithology Classification: Insights into Machine Learning Methods. SPE Annual Technical Conference and Exhibition,

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of Machine Learning*. MIT press.

Moiola, R. J., & Spencer, A. B. (1979). Differentiation of aeolian deposits by discriminant analysis. *US Geological Survey Professional Paper*(1052), 53.

Moiola, R. J., Spencer, A. B., & Weiser, D. (1974). Differentiation of modern sand bodies by linear discriminant analysis. *Gulf Coast Association of Geological Societies Transactions, 24*, 321-326.

Muhammed, D. D., Simon, N., Utley, J. E. P., Verhagen, I. T. E., Duller, R. A., Griffiths, J., Wooldridge, L. J., & Worden, R. H. (2022). Geochemistry of Sub-Depositional Environments in Estuarine Sediments: Development of an Approach to Predict Palaeo-Environments from Holocene Cores. *Geosciences, 12*(1), 23. https://www.mdpi.com/2076-3263/12/1/23

Purkait, B., & Das Majumdar, D. (2014, Jul). Distinguishing different sedimentary facies in a deltaic system. *Sedimentary Geology, 308*, 53-62. https://doi.org/10.1016/j.sedgeo.2014.05.001

R Core Team. (2016). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.* In https://www.R-project.org/.

Reineck, H.-E., & Siefert, W. (1980). Faktoren der Schlickbildung im Sahlenburger und Neuwerker Watt. *Die Küste, 35*(35), 26-51.

Rijn, J. N. v., & Hutter, F. (2018). *Hyperparameter Importance Across Datasets* Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, United Kingdom. https://doi.org/10.1145/3219819.3220058

Sevon, W. (1966). Distinction of New Zealand beach, dune, and river sands by their grain size distribution characteristics. *New Zealand Journal of Geology and Geophysics*, *9*, 212-223.

Simon, N., Worden, R. H., Muhammed, D. D., Utley, J. E. P., Verhagen, I. T. E., Griffiths, J., & Wooldridge, L. J. (2021, Jun). Sediment textural characteristics of the Ravenglass Estuary; Development of a method to predict palaeo sub-depositional environments from estuary core samples. *Sedimentary Geology*, *418*, 105906, Article 105906. https://doi.org/10.1016/j.sedgeo.2021.105906

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, *25*, 2960 - 2968.

Sokolova, M., & Lapalme, G. (2009, 2009/07/01/). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427-437. https://doi.org/https://doi.org/10.1016/j.ipm.2009.03.002

Sun, Z., Jiang, B., Li, X., Li, J., & Xiao, K. (2020). A Data-Driven Approach for Lithology Identification Based on Parameter-Optimized Ensemble Learning. *Energies*, *13*(15), 3903. https://www.mdpi.com/1996-1073/13/15/3903

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168-192. https://doi.org/10.1016/j.aci.2018.08.003

Vincent, P. (1998). Particle Size Differentiation of Some Coastal Sands: A Multinomial Logit Regression Approach. *Journal of Coastal Research*, *14*(1), 331-336. http://www.jstor.org/stable/4298781

Virolle, M., Brigaud, B., Bourillot, R., Fenies, H., Portier, E., Duteil, T., Nouet, J., Patrier, P., & Beaufort, D. (2019, Apr). Detrital clay grain coats in estuarine clastic deposits: origin and spatial distribution within a modern sedimentary system, the Gironde Estuary (south-west France). *Sedimentology*, *66*(3), 859-894. https://doi.org/10.1111/sed.12520

Visher, G. S. (1969). Grain size distributions and depositional processes. *Journal of Sedimentary Research*, *39*(3), 1074-1106. https://doi.org/10.1306/74d71d9d-2b21-11d7-8648000102c1865d

Walker, R. G. (1990). Facies modeling and sequence stratigraphy. *Journal of Sedimentary Research*, *60*(5), 777-786. https://doi.org/10.1306/212f926e-2b24-11d7-8648000102c1865d

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Wooldridge, L. J., Worden, R. H., Griffiths, J., Thompson, A., & Chung, P. (2017a). Biofilm origin of clay-coated sand grains. *Geology, 45*(10), 875-878.

Wooldridge, L. J., Worden, R. H., Griffiths, J., & Utley, J. E. P. (2017b, Apr). Clay-coated sand grains in petroleum reservoirs: Understanding their distribution via a modern analogue. *Journal of Sedimentary Research, 87*(4), 338-352. https://doi.org/10.2110/jsr.2017.20

Wooldridge, L. J., Worden, R. H., Griffiths, J., & Utley, J. E. P. (2018). The origin of clay-coated sand grains and sediment heterogeneity in tidal flats. *Sedimentary Geology, 373*, 191-209. https://doi.org/10.1016/j.sedgeo.2018.06.004

Wooldridge, L. J., Worden, R. H., Griffiths, J., & Utley, J. E. P. (2019a). Clay coat diversity in marginal marine sediments. *Sedimentology, 66*, 1118-1138. https://doi.org/10.1111/sed.12538

Wooldridge, L. J., Worden, R. H., Griffiths, J., & Utley, J. E. P. (2019b). How to quantify clay-coat grain coverage in modern and ancient sediments. *Journal of Sedimentary Research, 89*, 135-146.

Xie, Y., Zhu, C., Zhou, W., Li, Z., Liu, X., & Tu, M. (2018, Jan). Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances. *Journal of Petroleum Science and Engineering, 160*, 182-193. https://doi.org/10.1016/j.petrol.2017.10.028

Zhang, W., Wu, C., Zhong, H., Li, Y., & Wang, L. (2021, 2021/01/01/). Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geoscience Frontiers, 12*(1), 469-477. https://doi.org/https://doi.org/10.1016/j.gsf.2020.03.007

Zheng, D.-Y., & Wu, S.-X. (2021). Principal component analysis of textural characteristics of fluvio-lacustrine sandstones and controlling factors of sandstone textures. *Geological Magazine, 158*(10), 1847-1861. https://doi.org/10.1017/S0016756821000418

Zheng, D., Hou, M., Chen, A., Zhong, H., Qi, Z., Ren, Q., You, J., Wang, H., & Ma, C. (2022, 2022/08/01/). Application of machine learning in the identification of fluvial-lacustrine lithofacies from well logs: A case study from Sichuan Basin, China. *Journal of Petroleum Science and Engineering*, *215*, 110610. https://doi.org/https://doi.org/10.1016/j.petrol.2022.110610

Zhong, R., Johnson, R., & Chen, Z. (2020, 2020/03/01/). Generating pseudo density log from drilling and logging-while-drilling data using extreme gradient boosting (XGBoost). *International Journal of Coal Geology*, *220*, 103416. https://doi.org/https://doi.org/10.1016/j.coal.2020.103416

Zhou, J., Qiu, Y., Zhu, S., Armaghani, D. J., Khandelwal, M., & Mohamad, E. T. (2021, 2021/10/01/). Estimation of the TBM advance rate under hard rock conditions using XGBoost and Bayesian optimization. *Underground Space*, *6*(5), 506-515. https://doi.org/https://doi.org/10.1016/j.undsp.2020.05.008

Zubillaga, J. J. K., & Edwards, A. C. (2005). Grain size discrimination between sands of desert and coastal dunes from northwestern Mexico. *Revista Mexicana de Ciencias Geológicas*, *22*(3), 383-390.

Kavzoglu, T., & Teke, A. (2022, 2022/04/22). Advanced hyperparameter optimization for improved spatial prediction of shallow landslides using extreme gradient boosting (XGBoost). Bulletin of Engineering Geology and the Environment, 81(5), 201. https://doi.org/10.1007/s10064-022-02708-w