communications
earth & environment

ARTICLE

https://doi.org/10.1038/s43247-022-00631-2 OPEN

An automatic sediment-facies classification approach using machine learning and feature engineering

An-Sheng Lee[1,2]✉, Dirk Enters[3], Jyh-Jaan Steven Huang[4], Sofia Ya Hsuan Liou[2]✉ & Bernd Zolitschka[1]

The delineation of sediment facies provides essential background information for a broad range of investigations in geosciences but is often constrained in quality or quantity. Here we leverage improvements in machine learning and X-ray fluorescence core scanning to develop an improved approach to automatic sediment-facies classification. This approach was developed and tested on a regional-scale high-resolution elemental dataset from sediment cores covering various sediment facies typical for the southern North Sea tidal flat, Germany. We use a machine-learning-built classification model involving simple but powerful feature engineering to simulate the observational behavior of sedimentologists and find that approach has 78% accuracy, followed by error analysis. The model classifies the majority of sediment facies and also, importantly, highlights critical sections for further investigation. Research resources can thus be allocated more efficiently. We suggest that our approach could provide a generalizable blueprint that can be applied and adapted for the research question and data type at hand.

[1] University of Bremen, Institute of Geography, Bremen, Germany. [2] National Taiwan University, Department of Geosciences and Research Center for Future Earth, Taipei, Taiwan. [3] Lower Saxony Institute for Historical Coastal Research, Wilhelmshaven, Germany. [4] National Taiwan University, Institute of Oceanography, Taipei, Taiwan. ✉email: alee@uni-bremen.de; yhliou@ntu.edu.tw

Sediment facies is defined as an assemblage of sediments that records specific characteristics of a depositional environment. It hence provides fundamental sedimentological information in space and time, which builds the necessary geological background for studying past surface processes on Earth[1]. For instance, the change of sediment facies through time assists scientists in reconstructing the evolution of paleoenvironments[2–4]. Moreover, the knowledge of sediment facies makes available a proper selection of geochemical proxies for a better interpretation of paleoclimatic variations[5–8]. Exploring geological resources, such as for methane hydrates, oil, and gas, or offshore wind parks also requires regional investigations of sediment facies[9,10]. Basically, all research initiatives rely on the classification of sediment facies as a preliminary step.

Evolution in computer science and computing power made machine learning (ML) techniques available to various disciplines over the past two decades[11]. Research has started to include ML to achieve in-depth scientific findings previously constrained by conventional methods and data resolutions[12,13]. ML applications also have been introduced to the geoscience community, such as support vector machine (SVC), random forest (RF), and artificial neural networks[14–18]. However, applications are still at an early stage. The data coverage usually remains small, which can be expressed in either the amounts of data (<10 k), target classes (e.g., sediment facies and geochemical zones, <6), or features (e.g., measuring analytes, <10). For instance, input data are mostly geophysical measurements (e.g., seismic profiles and borehole logging data)[14,18–21] or laborious measurements (e.g., grain-size variations and quantitative element concentrations)[15,17], which have resolution limits (tens of centimeters to meters in scale). Target classes are thus often homogeneous sediments (e.g., unique tephra layers or sandstones)[16,20–22] instead of complex units having the sedimentary structure in an mm scale (e.g., laminated stream channel deposits). Furthermore, the ML models' decision commonly depends on individual data points rather than on comprehensive observations, including adjacent data points, which compare much better to the way how sedimentologists work and investigate. Insua et al.[19] determine that their single-point-observation ML models perform worse on composite carbonates than sand layers. All these factors limit the complexity and diversity of applications. The same is true for the fact that developing codes are rarely open-source.

Evolution in core scanning techniques (e.g., X-ray fluorescence (XRF) core scanner, computed tomography (CT), multi-sensor core logger (MSCL), and hyperspectral imaging (HSI)) provide the possibility of acquiring near-continuous (μm-scale resolution), non-destructive and rapid measurements covering both geophysical and geochemical data from natural archives[8,23–25]. A large quantity of sediment investigations is thus integrated with higher spatial resolution and much more measuring detail[26]. Therefore, data variety and size ascend to a new level, which helps resolve previously unachievable scientific questions (e.g., paleoclimatic variation and anthropogenic interaction with the natural environment)[27–29]. Among these scanning techniques, XRF core scanning is selected for this study as the primary technique to acquire measurements. It is evolved from a well-developed chemometric community[30], and its applications have been prosperous since 2000[8]. Its rapid measuring provides a wide range of elemental signals[31], which are adopted by the vast Earth and environmental research society (e.g., paleoceanography, paleolimnology, paleoseismology, pollution history[29,32–34]). Besides the chemical properties, XRF core scanning data offer proxies representing physical and biological properties, such as grain size[35] and diatom productivity[33]. Physical measurements, like CT and MSCL techniques, provide more convenient and economical ways to acquire data compared to XRF core scanning, e.g., no core opening is needed. However, they are not direct or sufficient information to describe the sediment composition content[21]. HSI has a high potential in acquiring high-resolution chemical and physical measurements useful for ML applications, but it can be more precise when coupled with XRF data[36]. XRF core scanning may give the most comprehensive and reliable information to simulate the sedimentologists' observational behavior, which includes multiple analyses (e.g., grain size, carbonate content, organic matter, and fossils). Currently, unsupervised ML applications with XRF data are common for data exploration, but not many developments in supervised ML[8,30,36].

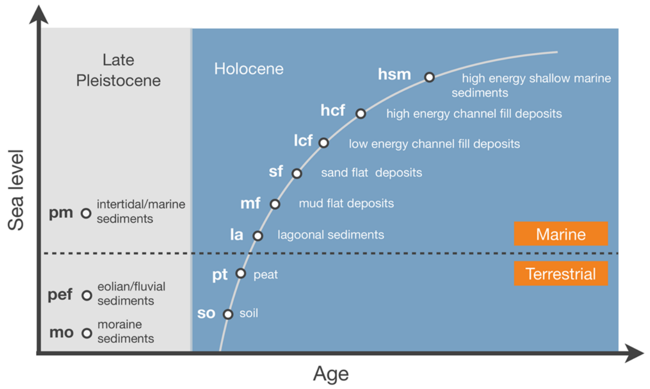
Sea level
Age
Late Pleistocene
Holocene
pm intertidal/marine sediments
pef eolian/fluvial sediments
mo moraine sediments
hsm high energy shallow marine sediments
hcf high energy channel fill deposits
lcf low energy channel fill deposits
sf sand flat deposits
mf mud flat deposits
la lagoonal sediments
pt peat
so soil
Marine
Terrestrial

Fig. 1 Sediment facies information. Abbreviation, age, and sedimentary environment for facies occurring in the study area, a tidal flat in the northern German Wadden Sea (adopted from Capperucci et al. (2022)[69]). Positions of facies in relation to sea level and age are only for illustration. For details, see Supplementary Table S1.

Our approach aims at enhancing the automatic sediment-facies classification by overcoming most restrictions. This is inherited from the successes of two evolutions: ML and core scanning techniques. High-resolution elemental profiles obtained by XRF core scanning were acquired under the interdisciplinary Wadden Sea Archive (WASA) project[26] from the UNESCO World Heritage-listed Wadden Sea in Northern Germany. This region belongs to a dynamic geological setting, where the environment varies due to the rapidly rising sea level during the Holocene from glacial (terrestrial) to shallow marine[2,37,38]. Thus, the studied sediments provide comprehensive geochemical signals and broad coverage (11 sediment facies, 53 core sections, 19k data points, for details: Fig. 1 and Methods), spanning the Late Pleistocene throughout the Holocene[37]. Instead of effortlessly applying ML algorithms, we tested different feature engineering methods, like data transformations and principal component analysis (PCA), to simulate the sedimentologists' observational behavior by extending the model's single-point to a multi-point analysis. Simple (logistic regression (LR)) and complex (kernel SVC, RF) ML algorithms were included to compare their feature engineering-integrated performance. Careful evaluations involving cross-validation (CV), tailor-designed measures, and error analysis were carried out during these steps to guarantee the generalization of the models.

The results indicate a noticeable benefit of feature engineering that empowers the use of simple ML algorithms, giving enhanced performance (78% accuracy) and applicability to a standard computer. In addition, the optimal model provides an outcome with a confidence level, which highlights critical parts of the sediment records to allocate more resources for analyzing and saving resources from the remaining sediments. The approach is expected as a blueprint to inspire further studies in Earth and environment communities regardless of data type, region, and objective. More importantly, this brings interdisciplinary

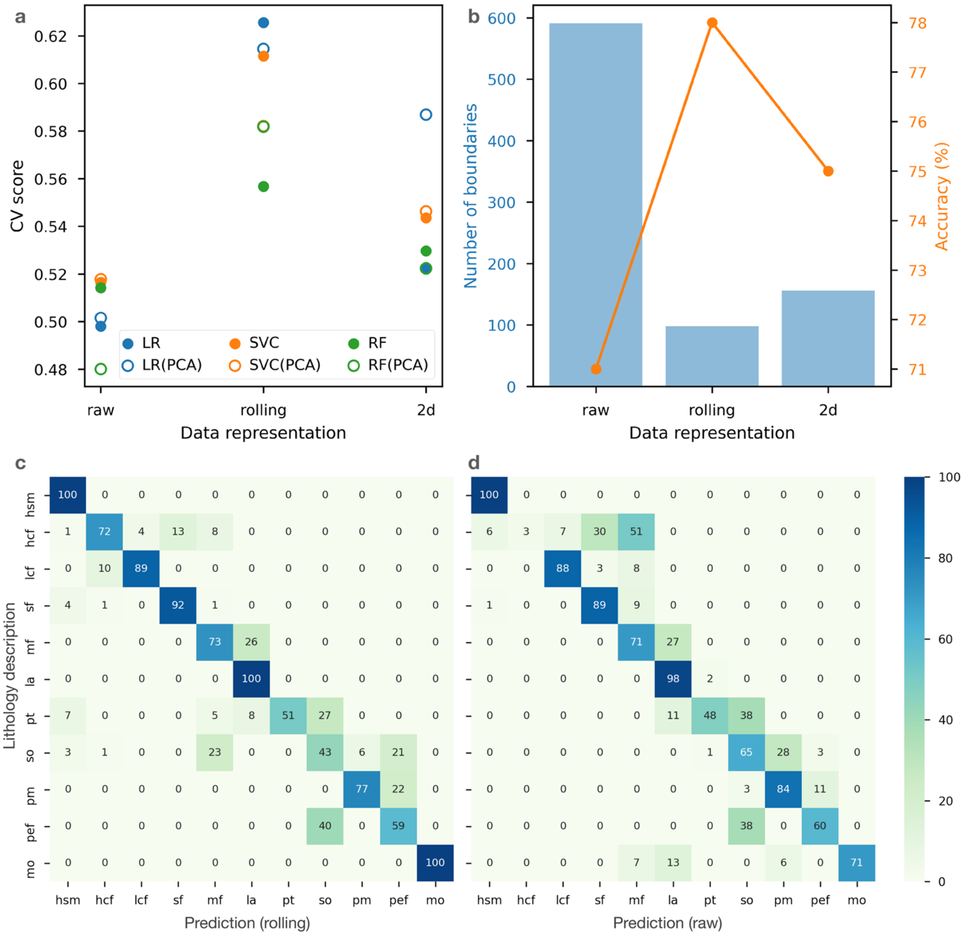

Fig. 2 Evaluations of built models. a The best cross-validation (CV) score of each combination of feature engineering (data representations and principal component analysis—PCA) and ML algorithms (logistic regression—LR, kernel support vector machine—SVC, random forest—RF classifiers). The CV score is relevant to the mean accuracy during CV iterations. **b** The fragmentation is quantified by the number of boundaries and accuracy of the optimal model built from each data representation in the test set. **c**, **d** The modified confusion matrices describe the performance of the optimal models built from (**c**) rolling data and (**d**) raw data when applied to the test set. The *y*-axis represents the facies from the lithological description by sedimentologists. The *x*-axis stands for the model-classified facies. The numbers represent the percentages of data in each row, i.e., recall in statistical terminology. For instance, at the top left of (**d**), only 3% of data points recognized as hcf by sedimentologists are correctly classified by the model, while 51% of these data points are misclassified as mf by the model. For abbreviations, refer to Fig. 1.

scientific research to a new era and scale of capability based on state-of-the-art technological innovations.

Results

Contribution to feature engineering. A series of models were built to classify sediments using the high-resolution elemental profiles automatically. The performances from each combination of feature engineering and ML algorithms were evaluated by the mean accuracy of five-fold CV scores. The output and visualization of the grid search are presented in Supplementary Figs. S3–5. The highest CV scores for each combination are illustrated in Fig. 2a. The CV scores increase after data representations, and the rolling representation results in the best scores. Only 5 out of 9 comparisons between using or not using PCA transformation indicate “including PCA” as the better feature engineering. Thus, there is no clear advantage of using a PCA transformation. After data representation, the complex ML algorithms (SVC and RF), which have the enhanced learning power of non-linearity, provide no noticeable advantage over the simple algorithm (LR).

Furthermore, the best models using each kind of data (Fig. 2a, raw data: SVC with PCA, rolling data: LR without PCA, 2D data: LR with PCA) were evaluated in the test set. Figure 2b shows that the model using rolling data provides the highest accuracy (78%) and fewest boundaries (98). Its boundary amount is about 20 times that sedimentologists gave (5), which is still noticeably lower than the model using raw data output (591). The model using 2D data does not perform as well as the one with rolling data but is still better than raw data. When considering the large size of 2D data (nine times rolling data), creating a burden for computing, the preference for the rolling data approach is confirmed.

The optimal model built on the best ML combination (LR learned from rolling data without PCA) indicates its increasing performance on laminated facies. When comparing the accuracies of channel fill deposits (especially hcf, Fig. 2c) to those of the model using raw data (Fig. 2d), the misclassification to sand and mud flats is remarkably reduced. This points out that the rolling representation describes the laminated characteristic of channel fill deposits well enough, comprising two kinds of homogeneous sublayers (sand and mud). The optimal model does not classify those sediments separately but together as one facies to recognize a specific depositional environment. The simple algorithm is thus allowed to surpass more complex algorithms to the benefit of

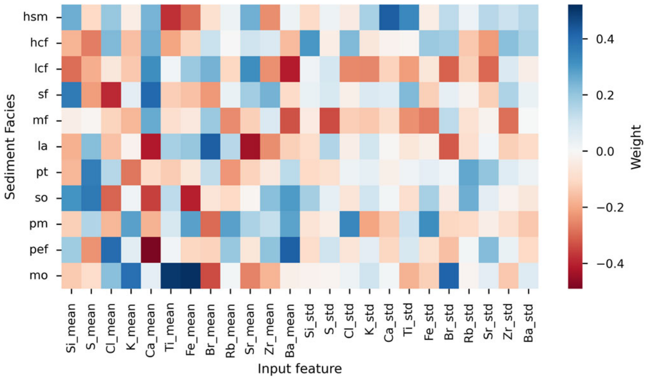

Fig. 3 The optimal model's discriminating behavior. Weights of features in the decision function of the optimal model built using LR from the rolling data without PCA transformation. The *x*-axis contains the features of rolling data (i.e., rolling mean and standard deviation of the raw elements). The *y*-axis is the facies (abbreviations refer to Fig. 1). The darker color, the more decisive that feature is in classifying to a specific facies. A negative coefficient indicates the negative influence of the feature.

feature engineering. This finding offers an efficient way of applying automatic sediment-facies classification by demanding less computing power due to less model complexity.

Decision function of the model. Owning to the capability of LR, our optimal model built using LR on the rolling data without PCA transformation gives weights describing which features (rolling mean and standard deviation of elements) are discriminating for each facies (i.e., decision function, Fig. 3). In the following expression, the elements refer to their rolling mean values if without "variance" specification, which represents the essence of standard deviation. Classifying as hsm has (1) substantial negative weights on Ti and Fe, (2) positive weights on Si, Ca, and Sr, and (3) positive weights on the variance of Ca and Ti. The channel fill deposits (hcf and lcf) positively rely on Ca but negatively on Si. To distinguish these two facies, hcf has an evident influence through Si variance, while lcf has a noticeable negative weight on Ba. Furthermore, the decision of lcf is less affected by the variances of elements compared to hcf. sf has substantial positive weights on Si and Ca but negative on Cl, and mf has a positive weight on Ca while negative weights on Rb and Ba. It has an overall negative weight on the variance of elements, especially on S. la has (1) strong negative weights on both Ca and Sr, (2) an apparent positive weight on Br but a negative weight on its variance. pt has a positive weight on S and a negative weight on K. It also has some weights on the variance of Rb and Sr., so has remarkable positive weights on Si, S, and Ba but negative weights on Cl, Ca and Fe. pm has (1) positive weights on K, Fe, Rb, and Ba, (2) negative weights on Ca and Br, and (3) positive weights on the variance of Cl and Fe. pef has noticeable positive weights on Cl and Ba but a negative weight on Ca. mo has the highest (1) positive weights on K, Ti, and Fe, (2) a negative weight on Br, and (3) positive weight on the variance of Br when compared to other facies.

Misclassification of the model. Besides the improvements that benefited from feature engineering, the optimal model has limitations causing misclassifications. Error analysis summarizes four main error-causing categories (Fig. 4a). The first category, with 63.6% occurrence frequency, describes the situation that the predictions could be correct, but this facies change is omitted to fit a general picture of environmental interpretation by sedimentologists. This is because the capture of composite characteristics is limited by the initially defined (fixed) window size during the rolling representation. For those laminations or sediment sections having a thickness >32 mm (window size: 17 data points), the optimal model cannot identify them as composite facies. These misclassifications are often considered minor sedimentary structures. For instance, minor low-energy channel fill deposits are misclassified to high-energy channel fill deposits in core Sections N31–1 (Fig. 4b), which might reveal a small-scale channel deepening. In fact, sedimentologists have a flexible observation window to identify facies, but our ML-based optimal model can only start from a local measurement and joins with nearby information. Therefore, these misclassifications should be considered as the model looks too detailed rather than classification problems.

Another error category is the transition boundary problem, having a 41.2% occurrence frequency (Fig. 4a). Figure 2c shows that the accuracies for peat, soil, Pleistocene eolian/fluvial deposits, and Pleistocene marine sediments are relatively low compared to the other facies. In most cases, the boundaries between these facies are gradual and/or indistinct. It is difficult for sedimentologists to draw a clear separation between them via macroscopic observation and elemental variation. For example, soils developed from Pleistocene sediments are similar in their elemental composition and cause confusion in the model. This cause of misclassification can be found in the boundaries of pt-so-pef (lower part of N71-4, Fig. 4b). The relevant transition boundary bias has also been mentioned in a pilot automation study nearby[21].

The third category having a 21.4% occurrence frequency (Fig. 4a), is related to the limitation of macroscopic observation by sedimentologists. Sedimentologists describe these misclassified sediments with doubts since they are visually similar to several sediment facies. For instance, in N18-2 (Fig. 4b), around 0.6 m depth, structureless sediments are described as eolian/fluvial deposits like their neighbors by sedimentologists. However, according to their elemental profiles, the optimal model classified them differently from the neighbors as Pleistocene marine sediments. This is reasonable because the Pleistocene sand flat sediments might be visually identical to the structureless eolian/fluvial deposits unless checking their diatom and geochemical data.

Subjective and opposing judgments also happen between sedimentologists. Our model provides a consistent judgment through its fixed and relatively small observation window. Only a few cases, like wood fragments, are misclassified as high-energy shallow marine sediment. But these are out of the model's ability. This material (3.7% occurrence frequency, Fig. 4a) is not in the facies list of training data for model building. The detailed error-analysis result and comparison results are listed in Supplementary Data 1 and Figs. S6 and 7.

Highlighting critical segments. Our optimal model estimates the confidence level of each classification. Figure 5a demonstrates an example of the probability distribution along the core depth of one selected core section. The probability values are dispersed along the facies if the model lacks confidence in its classification. In contrast, the high probability in specific facies stands for the model's confident decision. To more conveniently recognize the confidence level of our model, the maximum probability among facies for each data point is extracted (Fig. 5b). The higher the maximum probability value, the more confident the decision is.

The relation between the model's confidence level and its error rate is illustrated in Fig. 5c. The misclassifications (labeled as incorrect) mainly have a low maximum probability, suggesting that the more uncertain our model is, the more likely it is wrong. Empirically, the model's classification has a higher chance of

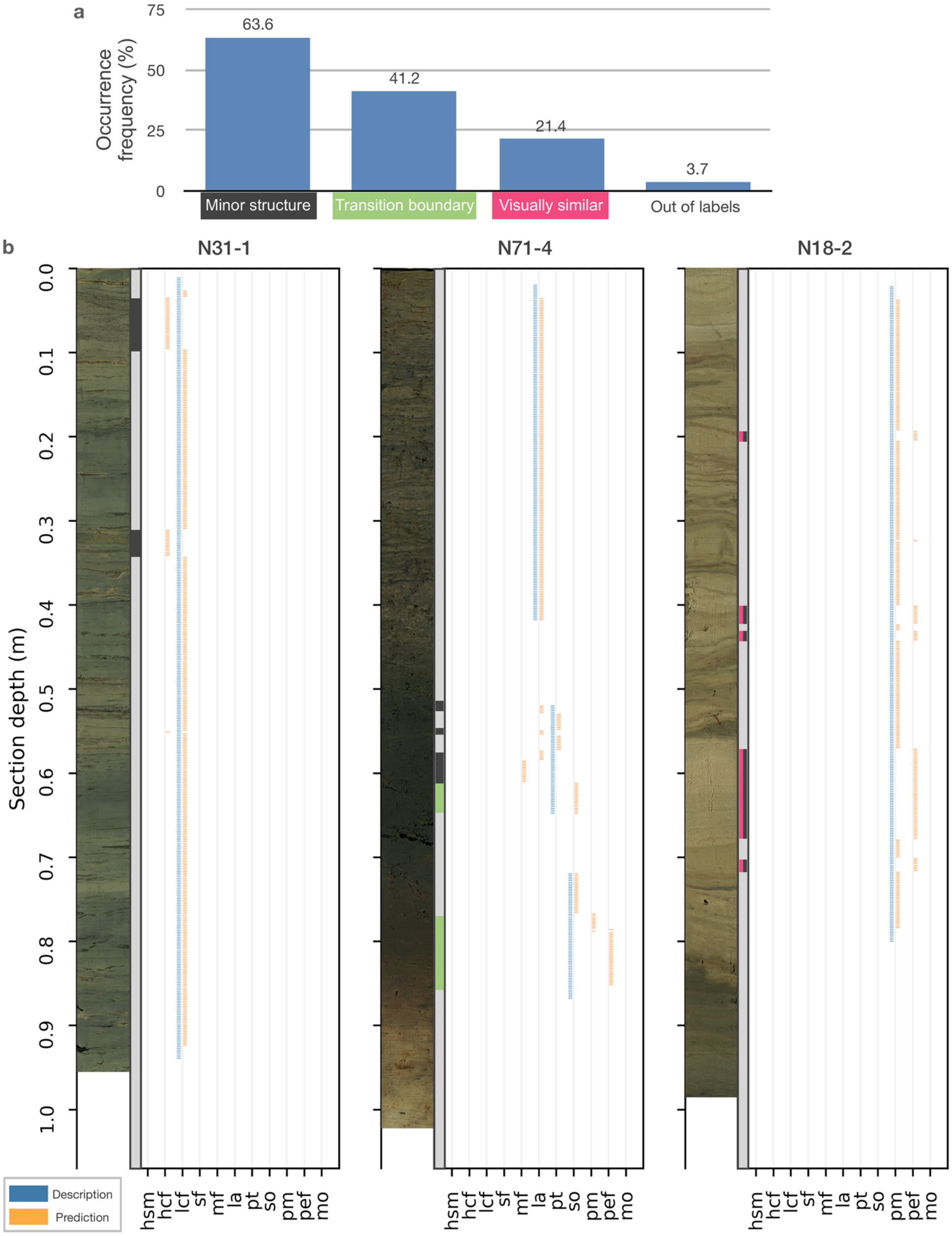

Fig. 4 Error analysis results from the test set. a Error categories and their frequency in misclassifications of the test set. The sum of the frequency is not 100% because categories may co-exist. **b** Comparison (three illustrating cores) between sedimentologists' descriptions (blue) and predictions (orange) of the optimal model built using LR from the rolling data without PCA transformation. The core images and error category labels (marked in the color code (**a**) are attached. For abbreviations, refer to Fig. 1.

being wrong when the maximum probability value is <0.3. Therefore, sedimentologists can focus on sediment sections with a low maximum probability to carry out further examinations, such as microscopic observations or quantitative geochemical analyses, while adopting the automatic classification done by the model if the maximum probability is >0.3. This provides a valuable outcome by highlighting the critical part where the expertise and experience of sedimentologists are needed for correct facies classification. When applying the model to the whole dataset of WASA (all 92 cores, 383 sections, 159 k data points), only 1% of the sediments are marked as critical, requiring further analysis.

Discussion

A simple but powerful recipe of feature engineering (centered log-ratio transformation and rolling representation) is proposed to fit the specific needs of sedimentological applications after testing several combinations of feature engineering methods. The best model does not need complex ML algorithms (SVC and RF), thus reducing computation time and model size. A possible reason explaining the usefulness of rolling representation is its combination of measurements (i.e., chemical, physical, and biological-related XRF data) and heterogeneities from the sample and adjacent sediments. This data representation moves from a single-point observation to a multi-point (window-sized) observation, which simulates human observation by considering the context of sedimentary features to define the sediment facies. The rolling mean draws an overall elemental fingerprint of the sediments, while the rolling standard deviation describes heterogeneity within this fingerprint.

By interpreting the optimal model's decision function (Fig. 3), we can realize the model's classification behavior and inspect its

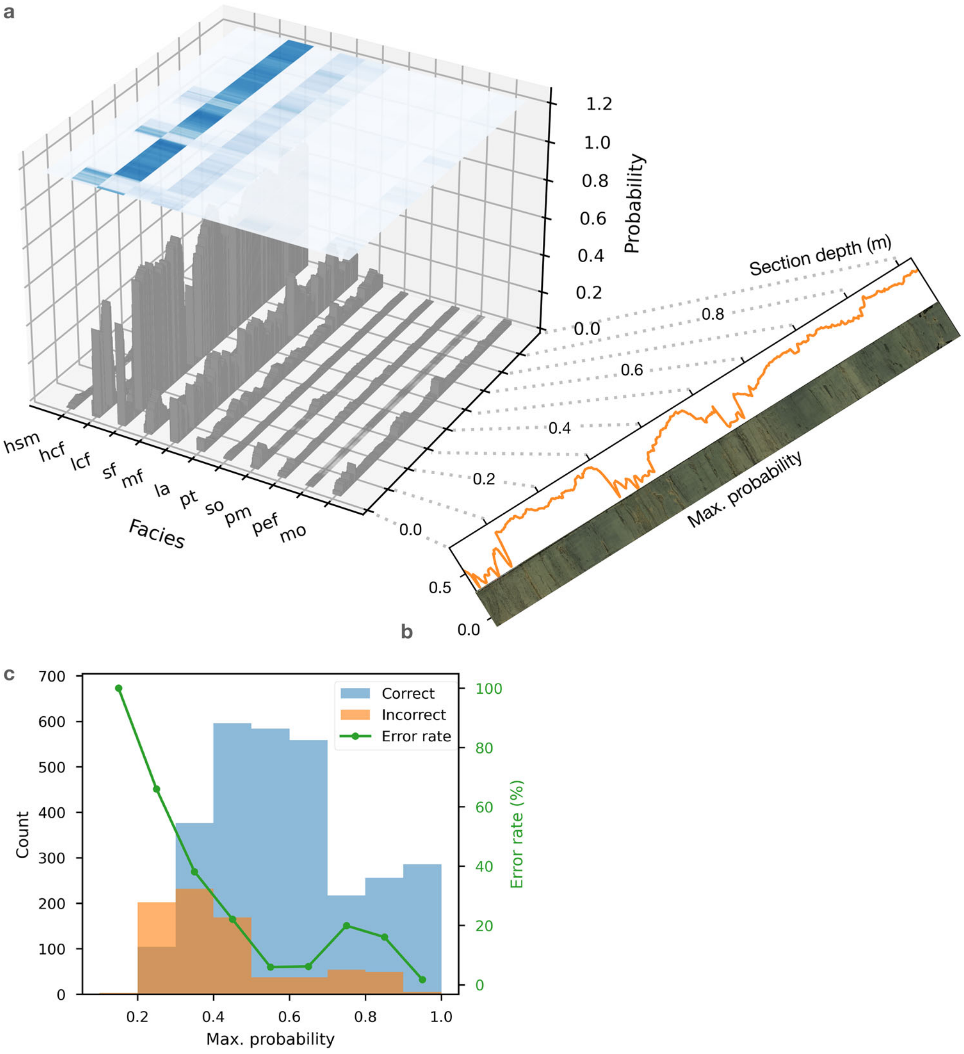

Fig. 5 Attributes of the optimal model's confidence level. a Probability along core Sections N31–1. The probability of each facies in each data point is visualized in bars and projected to a heatmap above (the deeper the blue, the higher the probability). **b** The maximum probability for each data point is plotted along with the core image. For abbreviations, refer to Fig. 1. **c** Bi-plots of (1) distribution of the maximum probability in the groups of correct and incorrect classifications and (2) error rate for each interval of the maximum probability in the test set.

justifiability. The weights in classifying hsm reflect that the model characterizes the alternating coarse sand (Si) and shell fragments (Ca, Sr) as its standard and has a dominant marine influence[39–41]. The model takes the abundance of shell fragments and relatively few sand grains as the key to discriminating channel fill deposits. Notably, the model recognizes more contrasting sediment properties in hcf than lcf by considering the variances of individual elements. The negative contribution of Ba, which correspondingly happens to mf, may show the lack of this element in finer sediments (clay minerals). Having a high contribution to the elements related to coarse grain (Si) and shell fragment (Ca) and a negative contribution to seawater (Cl) describes the composition of sf and its distance to the sea correctly[39–41]. The model depicts the homogeneous characteristic of mf by giving a negative contribution to the variances of individual elements. la is characteristic of a lower contribution from shell fragments, but more homogeneous organic matter (Br) is recognized by the model[34]. Interestingly, the model uses the proxy for pyrite (S), less contribution of mineralogical sediments (K), and the slightly heterogeneous composition based on the variance to differentiate pt from la, which is also organic-rich[7,42]. The weights for distinguishing paleosoils can be related to the leaching (Ca, Fe) process[43–45], but also the transition boundaries to adjacent sediments, which creates mixed signals and usually leads to bias (e.g., the influence of above peat (S)). Although the sedimentation process of pm is relevant to those Holocene shallow marine sediments, our model learns a major difference between them: lacking carbonate and organic matter due to long degradation. Meanwhile, the abundance of K and Ba is a vital differentiation possibly caused by different sediment provenances[46]. The distinct characteristics of pef are carbonate-free and structureless, which are well described by the model's weights. However, the reason for the model considering Cl and Ba is unclear. To classify mo from other facies, the model maps its function in having the highest weights for mineralogical elements (K, Ti, Fe) and a negative contribution of organic matter (Br). Most of the weights in the decision function are sensible. A few of them do not have an appropriate explanation, possibly due to the bias caused by rolling transformation.

The comprehensive evaluation of the model is estimated by using robust CV in the training set as well as boundary amount and balanced accuracy in the test set. Based on a performance of 78% accuracy and the result of detailed error analysis, most misclassifications are summarized as (1) minor structures that were ignored by sedimentologists, (2) confusion caused by transition boundaries, and (3) sediments visually identical to several facies (Fig. 4). These error categories are related to the subjective decision by sedimentologists and thus are described as unavoidable bias[47]. They can also be considered as opinion variations among sedimentologists, not actual errors. Therefore, it is important for sedimentologists to identify these arguable sediment sections (as quantified by the confidence level, Fig. 5a) from the entire sediment record to concentrate efforts and discussions.

As mentioned for the model's limitations, our approach has not yet fully simulated the sedimentologist's behavior. With a flexible observational window and a bigger picture of environmental interpretations in mind, the complex or customized architecture of neural network algorithms, like long short-term memory, may cause improvements. Also, more sediment facies can be considered, which should boost the success of automation. Numerous ML applications use images containing pattern recognition, segmentation, and further well-built deep learning architecture (e.g.,[11,48,49]). Replacing XRF data with images may be a worthy attempt because it gives higher resolution at a lower measurement cost. Yet, its nature of losing some information that XRF data has needs to be aware.

Our study provides not only a model to automatically classify sediments from the Wadden Sea in Germany, but also can be considered as a methodological blueprint for projects, regardless of the type of data and the region investigated. Compared to other studies developing an automatic classification of geological records[15,17,18,21], our approach has elevated the application to a new level, which has (1) input data with comprehensive information that is widely accepted in the Earth and environmental sciences and (2) broader data coverage and facies variability. Scientifically, a large-scale scientific investigation, such as the whole Wadden Sea area (>1500 km^2), needs a less time- and labor-consuming research approach to be accomplished[21]. In industrial terms, cost-efficiency and the capability of managing large-scale exploration data are in demand. Since our model can automatically classify the relatively simple but often vast sediment sections, experts can deal with critical sections and focused analyses. Thus, the resources (time, personnel, material) can be redistributed more efficiently. Our future toward a standard operative procedure of automatically coring, scanning, classifying, and digitally archiving sediments requires further systematic scientific and commercial developments. Hopefully, our advances will assist in exploring new perspectives on Earth and environmental sciences. Furthermore, this evolution shall facilitate the preservation of geoscientific knowledge since investigations will be conducted digitally and promote the findability, accessibility, interoperability, and reusability principles[50] (especially reusability).

Methods

Data acquisition. This study was developed under the framework of the interdisciplinary WASA project[26]. Ninety-two sediment cores (length: up to 6 m, diameter: 8 and 10 cm) were recovered from tidal flats, channels, and offshore around the island of Norderney. A team of sedimentologists carried out lithological descriptions and sediment-facies interpretation through macroscopic observation[37]. High-resolution elemental profiles were acquired from a COX Itrax-XRF core scanner at the GEOPOLAR lab, University of Bremen, scanned at a fixed setting, and subsequently processed by the Q-spec software (version 2015; COX® Analytical Systems). Twelve elements (Si, S, Cl, K, Ca, Ti, Fe, Br, Rb, Sr, Zr, Ba) were chosen based on signal reliability. All data and information were adopted from a previous study[42] that compiled the geochemical and geophysical measurements of the sediment cores. Fifty-three representative sediment core sections (length: <1.2 m, locations: Supplementary Fig. S1a, with 19,823 data points) covering 11 sediment facies (Fig. 1) were selected for this approach in order to constrain computing time. Data points representing cracks and rough or uneven sediment surfaces were excluded. The facies labels extracted from lithological descriptions and elemental profiles were aligned according to their depth.

Feature engineering. The developing scheme of this study is demonstrated in Supplementary Fig. S2a. The dataset was randomly split into training and test sets for a robust evaluation. The test set contains one section for each facies while the remaining sections were used for training. A series of feature engineering methods (i.e., data transformations in ML terminology) were deployed to enhance the performance of the models built by supervised ML algorithms. Then, combinations of these methods were tested to find the most useful one. The steps are the following.

Elemental data were normalized by the geometric mean in each data point to eliminate the variance caused by the machine and the sediment itself, such as XRF tube aging, water content, and grain size, to achieve a better prediction. In addition, the logarithm was applied to free the normalized data from asymmetry and closed-sum effects[51,52]. Together, this data treatment is called centered log-ratio transformation, which is commonly applied to compositional data[53].

To capture the composite character of sediment facies, the neighboring information of each data point needs to be included in the data point. We propose two data representations for comparison. The first is rolling representation, where the dataset is represented by a centered moving mean and its standard deviation (Supplementary Fig. S2b) for each element. The moving window size was set to 17 data points since the step-size resolution of the Itrax XRF core scanner is 2 mm, and the thickness of beddings is predominantly <1 cm for studied sediments[42]. The second is called 2D representation. It collects adjacent data points as a chunk of data (17 data points * 12 elements) and raveled them to an array of new dimensions (Supplementary Fig. S2c). This approach is common in image analysis, which transforms a 2D pixel data matrix into a 1D array. These representations were implemented for each core section individually to prevent the model building from data snooping[54].

In the next step, the represented dataset was standardized to zero mean and unit standard deviation, which is essential for some ML algorithms. PCA with correlation matrix and whitened settings was included in the workflow to discuss its need in feature engineering. Yet, the standardization was not applied to the dataset for the combination of RF classifier without PCA because RF is not sensitive to the variable's scale difference.

Model building and evaluation. After feature engineering, three algorithms were applied to learn the dataset and to build models to classify sediments into facies by analyzing elemental profiles automatically. LR classifier using 'lbfgs' as solver[55] and L2 regularization was selected from linear algorithms. LR learns from the data to a model with a linear combination of weights derived from the maximum likelihood convergence. It is a classic soft classifier that produces probability as an outcome in dealing with multi-class problems[56]. Hence, it quantifies the uncertainty of predictions rather than just classification results. Kernel SVC and RF classifiers were selected from sophisticated non-linear algorithms. The kernel technique of SVC allows for exploring data relations in the infinite space, which includes element ratios. Furthermore, SVC has the ability to tolerate noise due to its soft-margin policy when finding the hyperplane in classification[57]. RF is an aggregation of decision trees that enforces its non-linear learning capability and decreases the overfitting potential by the nature of randomization (both samples and features). Each decision tree is constructed by a series of decision-making to separate samples into the desired classification, which simulates the human's intuitive behavior. In addition, the decision tree is free from the scale of data, so there is no need for normalization prior to the implementation, which is more convenient and leads to a difference in the workflow (Fig. S2a)[58,59]. These three algorithms are well-developed in the ML community and commonly adopted by other disciplines[59,60]. The algorithms were implemented with the balanced class (i.e., facies) weights[61] to deal with our imbalanced facies distribution (Supplementary Fig. S1b).

Prior to the final evaluation, an exhaustive search for parameters using five-fold CV was utilized to (1) assess the usefulness of data representation, (2) evaluate the need of using PCA, and (3) fine-tune parameters (LR: C, SVC: C and γ, RF: n_estimators and max_depth) for building models. In brief, those parameters control the regularization strength of algorithms to achieve a better generalization[59]. C for LR refers to an inverse of regularization strength that penalizes the cost function based on L2 regularization[54,60]. C for SVC stands for the penalty for SVC's hyperplane margin in the cost function. A higher C characterizes a harder margin that tolerates fewer misclassifications (potential noise) according to its trained hyperplane. γ for SVC is a parameter in the Gaussian kernel (also known as Radio Basis Function) that defines how effective a single training sample is. These two parameters control the regularization but with a different approach. Both C (for LR and SVC) and γ should be positive floats[54,60]. We started the searching range from the default 1.0 (C) and 1/n_features (γ) in the log10 interval. n_estimators refer to the number of decision trees composing the forest (RF aggregation). A higher number of trees provides a higher randomization level and thus creates higher regularization. max_depth restricts the size of each decision tree to avoid overfitting[54,60]. We searched n_estimators from 100 to 5000 with a max_depth from 3 to 15 to find the optimal performance while compromising the computer memory consumption. The details of parameter searching are listed in Supplementary Figs. S3–5.

During each iteration of CV, the training set was split into two sets (second training and validation sets) following the splitting strategy. The 2nd training set was learned by a pipeline composed of feature engineering and algorithm with specific parameters to build a model. The model's performance was evaluated by its accuracy for the validation set. Instead of using simple accuracy, which calculates the overall percentages of correct predictions, we used balanced accuracy, a macro-average of recall scores per class or, equivalently, the mean of accuracies from 11 facies[62,63] for evaluation. This helps with the search for optimal parameters that build models predicting all facies fairly well rather than predicting the dominant facies only. A mean validation accuracy (CV score) was calculated after iterations, which is more robust in facing data heterogeneity. In total, 18 CV scores were collected from the best CV scores of each combination (3 kinds of data, with or without PCA, 3 algorithms).

Once the optimal parameters were decided based on CV results, these algorithms learned the whole training set to build optimal models. The test set was then used for evaluating the models' performances. The performance was described by balanced accuracy, demonstrated in a modified confusion matrix, and the number of boundaries, describing the fragmentation problem of predicted facies.

Error analysis was also carried out on the test set to investigate the underlying causes of misclassifications and consequently prioritize possible improvements[47]. Misclassifications were thoroughly checked via both elemental profiles and images to summarize error categories. Each misclassification was given one or multiple error categories. In the end, the occurrence amount of each category was divided by the total amount of misclassification as frequency, i.e., the category's contribution.

All computations and visualizations were conducted using the SciPy ecosystem in Python[60,64–68]. The function of providing model probability was adopted from the well-developed Scikit-learn package[60].

Data availability

The data is published in accordance with and under the license "CC-BY: Creative Commons Attribution 4.0 International": Lee, An-Sheng; Enters, Dirk; Zolitschka, Bernd (2022): XRF down-core scanning profiles and sediment facies of sediments from the East Frisian Wadden Sea, Germany. PANGAEA, https://doi.org/10.1594/PANGAEA.946251.

Code availability

All computer codes and instructions are open-access on https://github.com/dispink/WASA_faciesXRF.

Received: 4 May 2022; Accepted: 16 November 2022;
Published online: 26 November 2022

References

1. Reinick, H. E. & Wunderlich, F. Classification and origin of flaser and lenticular bedding. *Sedimentology* **11**, 99–104 (1968).
2. Streif, H. Sedimentary record of Pleistocene and Holocene marine inundations along the North Sea coast of Lower Saxony, Germany. *Quat. Int.* **112**, 3–28 (2004).
3. Kropelin, S. et al. Climate-driven ecosystem succession in the Sahara: the past 6000 years. *Science* **320**, 765–768 (2008).
4. Karle, M., Bungenstock, F. & Wehrmann, A. Holocene coastal landscape development in response to rising sea level in the Central Wadden Sea coastal region. *Neth. J. Geosci.* **100**, e12 (2021).
5. Sheldon, N. D. & Tabor, N. J. Quantitative paleoenvironmental and paleoclimatic reconstruction using paleosols. *Earth Sci. Rev.* **95**, 1–52 (2009).
6. Davies, S. J., Lamb, H. F. & Roberts, S. J. Micro-XRF core scanning in palaeolimnology: recent developments. in *Micro-XRF Studies of Sediment Cores: Applications of a Non-destructive Tool for the Environmental Sciences* (ed Croudace, I. W., Rothwell, R. G.) 189–226 (Springer Netherlands, 2015).
7. Rothwell, R. G. & Croudace, I. W. Twenty years of XRF core scanning marine sediments: what do geochemical proxies tell us? in *Micro-XRF Studies of Sediment Cores: Applications of a Non-Destructive Tool for the Environmental Sciences* (ed Croudace, I. W., Rothwell, R. G.) 25–102 (Springer, 2015).
8. Croudace, I. W., Löwemark, L., Tjallingii, R. & Zolitschka, B. Current perspectives on the capabilities of high resolution XRF core scanners. *Quat. Int.* **514**, 5–15 (2019).
9. Fujii, T. et al. Geological setting and characterization of a methane hydrate reservoir distributed at the first offshore production test site on the Daini-Atsumi Knoll in the eastern Nankai Trough, Japan. *Mar. Pet. Geol.* **66**, 310–322 (2015).
10. Coughlan, M., Long, M. & Doherty, P. Geological and geotechnical constraints in the Irish Sea for offshore renewable energy. *J. Maps* **16**, 420–431 (2020).
11. Jordan, M. I. & Mitchell, T. M. Machine learning: trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).
12. Wu, S. et al. Artificial intelligence reveals environmental constraints on colour diversity in insects. *Nat. Commun.* **10**, 4554 (2019).
13. Yu, Y. et al. Machine learning–based observation-constrained projections reveal elevated global socioeconomic risks from wildfire. *Nat. Commun.* **13**, 1–11 (2022).
14. Ai, X., Wang, H. & Sun, B. Automatic identification of sedimentary facies based on a support vector machine in the Aryskum Graben, Kazakhstan. *Appl. Sci.* **9**, 4489 (2019).
15. Bolandi, V., Kadkhodaie, A. & Farzi, R. Analyzing organic richness of source rocks from well log data by using SVM and ANN classifiers: a case study from the Kazhdumi formation, the Persian Gulf basin, offshore Iran. *J. Pet. Sci. Eng.* **151**, 224–234 (2017).
16. Bolton, M. S. M. et al. Machine learning classifiers for attributing tephra to source volcanoes: an evaluation of methods for Alaska tephras. *J. Quat. Sci.* **35**, 81–92 (2020).
17. Kuwatani, T. et al. Machine-learning techniques for geochemical discrimination of 2011 Tohoku tsunami deposits. *Sci. Rep.* **4**, 7044 (2014).
18. Wrona, T., Pan, I., Gawthorpe, R. L. & Fossen, H. Seismic facies analysis using machine learning. *Geophysics* **83**, 83–95 (2018).
19. Insua, T. L., Hamel, L., Moran, K., Anderson, L. M. & Webster, J. M. Advanced classification of carbonate sediments based on physical properties. *Sedimentology* **62**, 590–606 (2015).
20. Benaouda, D., Wadge, G., Whitmarsh, R. B., Rothwell, R. G. & MacLeod, C. Inferring the lithology of borehole rocks by applying neural network classifiers to downhole logs: an example from the Ocean Drilling Program. *Geophys. J. Int.* **136**, 477–491 (1999).
21. Hadler, H. et al. Automated facies identification by Direct Push-based sensing methods (CPT, HPT) and multivariate linear discriminant analysis to decipher geomorphological changes and storm surge impact on a medieval coastal landscape. *Earth Surf. Process. Landf.* **46**, 3228–3251 (2021).
22. Basu, T. et al. Automated facies estimation from integration of core, petrophysical logs, and borehole images. in *AAPG Annual Meeting* 1–7 (2002).
23. Ross, P. S., Bourke, A. & Fresia, B. A multi-sensor logger for rock cores: methodology and preliminary results from the Matagami mining camp, Canada. *Ore Geol. Rev.* **53**, 93–111 (2013).
24. Cnudde, V. & Boone, M. N. High-resolution X-ray computed tomography in geosciences: a review of the current technology and applications. *Earth-Sci. Rev.* **123**, 1–17 (2013).
25. Jacq, K. et al. Theoretical principles and perspectives of hyperspectral imaging applied to sediment core. *Anal. Quat.* **5**, 28 (2022).
26. Bittmann, F., Bungenstock, F. & Wehrmann, A. Drowned palaeo-landscapes: archaeological and geoscientific research at the southern North Sea coast. *Neth. J. Geosci.* **101**, e3 (2022).
27. Martin-Puertas, C., Tjallingii, R., Bloemsma, M. & Brauer, A. Varved sediment responses to early Holocene climate and environmental changes in Lake Meerfelder Maar (Germany) obtained from multivariate analyses of micro X-ray fluorescence core scanning data. *J. Quat. Sci.* **32**, 427–436 (2017).
28. Lintern, A. et al. Sediment cores as archives of historical changes in floodplain lake hydrology. *Sci. Total Environ.* **544**, 1008–1019 (2016).
29. Miller, H. et al. A 500 year sediment lake record of anthropogenic and natural inputs to Windermere (English Lake District) using double-spike lead isotopes, radiochronology, and sediment microanalysis. *Environ. Sci. Technol.* **48**, 7254–7263 (2014).
30. Panchuk, V., Yaroshenko, I., Legin, A., Semenov, V. & Kirsanov, D. Application of chemometric methods to XRF-data—a tutorial review. *Analytica Chim. Acta* **1040**, 19–32 (2018).
31. Croudace, I. W., Rindby, A. & Rothwell, R. G. ITRAX: description and evaluation of a new multi-function X-ray core scanner. *Geol. Soc. Lond. Spec. Publ.* **267**, 51–63 (2006).
32. Schwestermann, T. et al. Multivariate statistical and multiproxy constraints on earthquake-triggered sediment remobilization processes in the Central Japan Trench. *Geochem. Geophys. Geosyst.* **21**, e2019GC008861 (2020).
33. Zolitschka, B., Lee, A.-S., Bermúdez, D. P. & Giesecke, T. Environmental variability at the margin of the South American monsoon system recorded by a high-resolution sediment record from Lagoa Dourada (South Brazil). *Quat. Sci. Rev.* **272**, 107204 (2021).
34. Ziegler, M., Jilbert, T., de Lange, G. J., Lourens, L. J. & Reichart, G. J. Bromine counts from XRF scanning as an estimate of the marine organic carbon content of sediment cores. *Geochem. Geophys. Geosyst.* **9**, Q05009 (2008).
35. Bloemsma, M. R. et al. Modelling the joint variability of grain size and chemical composition in sediments. *Sediment. Geol.* **280**, 135–148 (2012).
36. Rapuc, W. et al. XRF and hyperspectral analyses as an automatic way to detect flood events in sediment cores. *Sediment. Geol.* **409**, 105776 (2020).
37. Schaumann, R. M. et al. The Middle Pleistocene to early Holocene subsurface geology of the Norderney tidal basin: new insights from core data and high-resolution sub-bottom profiling (Central Wadden Sea, southern North Sea). *Neth. J. Geosci.* **100**, e15 (2021).
38. Schlütz, F., Enters, D. & Bittmann, F. From dust till drowned: the Holocene landscape development at Norderney, East Frisian Islands. *Neth. J. Geosci.* **100**, e7 (2021).

39. Bahr, A., Lamy, F., Arz, H., Kuhlmann, H. & Wefer, G. Late glacial to Holocene climate and sedimentation history in the NW Black Sea. *Mar. Geol.* **214**, 309–322 (2005).
40. Piva, A. et al. Climatic cycles as expressed in sediments of the PROMESS1 borehole PRAD1-2, central Adriatic, for the last 370 ka: 1. Integrated stratigraphy. *Geochem. Geophys. Geosyst.* **9**, Q01R01 (2008).
41. Rothwell, R. G., Hoogakker, B., Thomson, J., Croudace, I. W. & Frenz, M. Turbidite emplacement on the southern Balearic Abyssal Plain (western Mediterranean Sea) during Marine Isotope Stages 1–3: an application of ITRAX XRF scanning of sediment cores to lithostratigraphic analysis. *Geol. Soc. Lond. Spec. Publ.* **267**, 79–98 (2006).
42. Lee, A.-S., Enters, D., Titschack, J. & Zolitschka, B. Facies characterisation of sediments from the East Frisian Wadden Sea (Germany): new insights from down-core scanning techniques. *Neth. J. Geosci.* **100**, e8 (2021).
43. Reineck, H.-E. & Singh, I. B. *Depositional Sedimentary Environments: With Reference to Terrigenous Clastics*. (Springer Science & Business Media, 2012).
44. Fischer, P. et al. Formation and geochronology of last interglacial to lower Weichselian loess/palaeosol sequences—case studies from the Lower Rhine Embayment. *Ger. EG Quat. Sci. J.* **61**, 48–63 (2012).
45. Kabata-Pendias, A. *Trace Elements in Soils and Plants*. (CRC press, 2000).
46. Diekmann, B. et al. Detrital sediment supply in the southern Okinawa Trough and its relation to sea-level and Kuroshio dynamics during the late Quaternary. *Mar. Geol.* **255**, 83–95 (2008).
47. Ng, A. *Machine Learning Yearning*. (deeplearning.ai, 2018).
48. Bankole, S. A., Buckman, J., Stow, D. & Lever, H. Automated image analysis of mud and mudrock microstructure and characteristics of hemipelagic sediments: IODP expedition 339. *J. Earth Sci.* **30**, 407–421 (2019).
49. Fabijańska, A., Feder, A. & Ridge, J. DeepVarveNet: automatic detection of glacial varves with deep neural networks. *Comput. Geosci.* **144**, 104584 (2020).
50. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
51. Weltje, G. J., et al. Prediction of geochemical composition from XRF core scanner data: a new multivariate approach including automatic selection of calibration samples and quantification of uncertainties. in *Micro-XRF Studies of Sediment Cores: Applications of a Non-destructive Tool for the Environmental Sciences* (eds Croudace, I. W., Rothwell, R. G.) 507-534 (Springer Netherlands, 2015).
52. Pawlowsky-Glahn, V. & Egozcue, J. J. Compositional data and their analysis: an introduction. *Geol. Soc. Spec. Publ.* **264**, 1–10 (2006).
53. Aitchison, J. The statistical analysis of compositional data. *J. R. Stat. Soc.: Ser. B (Methodol.)* **44**, 139–160 (1982).
54. Abu-Mostafa, Y. S., Magdon-Ismail, M. & Lin, H.-T. *Learning From Data: A Short Course*. (AMLBook, 2012).
55. Morales, J. L. & Nocedal, J. Remark on "algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization". *ACM Trans. Math. Softw.* **38**, Article 7 (2011).
56. Tu, J. V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J. Clin. Epidemiol.* **49**, 1225–1231 (1996).
57. Chang, C.-C. & Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2**, 1–27 (2011).
58. Breiman, L. Random forests. in *Machine Learning* 5-32 (Springer, 2001).
59. Müller, A. C., Guido, S. *Introduction to Machine Learning with Python: A Guide for Data Scientists*, 1 edn. (O'Reilly Media, 2016).
60. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
61. King, G. & Zeng, L. Logistic regression in rare events data. *Political Anal.* **9**, 137–163 (2001).
62. Brodersen, K. H., Ong, C. S., Stephan, K. E. & Buhmann, J. M. The balanced accuracy and its posterior distribution. In *International Conference on Pattern Recognition* 3121–3124 (2010).
63. Kelleher, J. D., Mac Namee, B. & D'arcy, A. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. (MIT press, 2020).
64. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 99–104 (2007).
65. McKinney, W. Data Structures for Statistical Computing in Python. in *The 9th Python in Science Conference* 56–61 (2010).
66. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
67. Millman, K. J. & Aivazis, M. Python for scientists and engineers. *Comput. Sci. Eng.* **13**, 9–12 (2011).
68. Virtanen, P. et al. SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **17**, 261–272 (2020).
69. Capperucci, R. M. et al. The WASA core catalogue of Late Quaternary depositional sequences in the central Wadden Sea-a manual for the core repository. *Netherl. J. Geosci.* **101**, e5 (2022).

Acknowledgements

The present study is jointly funded by (1) the Wadden Sea Archive (WASA) project funded by the 'Niedersächsisches Vorab' of the VolkswagenStiftung within the funding initiative 'Küsten und Meeresforschung in Niedersachsen' of the Ministry of Science and Culture of Lower Saxony, Germany (project VW ZN3197), and (2) the Ministry of Science and Technology of Taiwan (project number: MOST 110-2116-M-002-023). We appreciate the WASA project team that jointly carried out the core recovery and macroscopic sediment description. A.S.L. thanks the National Taiwan University (NTU) Research Center for Future Earth from the Featured Areas Research Center Program within the Higher Education Sprout Project framework funded by the Ministry of Education of Taiwan. A.S.L. appreciates the financial support from the 2022 Ørsted's Green Energy Scholarship Program. We acknowledge Ben Marzeion and Timo Rothenpieler (Klimageographie, University of Bremen) for generously providing high-performance computing power.

Author contributions

A.S.L.: conceptualization, methodology, investigation, data curation, writing—original draft. D.E.: conceptualization, methodology, writing—review and editing. J.J.S.H.: writing—review and editing. B.Z. and S.Y.H.L.: supervision, resources, funding acquisition, writing—review and editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43247-022-00631-2.

Correspondence and requests for materials should be addressed to An-Sheng Lee or Sofia Ya Hsuan Liou.

Peer review information *Communications Earth & Environment* thanks Man-Yin Tsang, Aqsa Anees, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Joe Aslin. Peer reviewer reports are available.