



Forecasting closures on shellfish farms using machine learning

Fearghal O'Donncha¹ · Albert Akhriev¹ · Bruno Fragoso^{2,3} · John Icely^{2,3}

Received: 5 April 2023 / Accepted: 16 February 2024 / Published online: 7 March 2024
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2024

Abstract

Biotoxins and harmful algal blooms (HABs) are damaging to aquaculture operations. Occurrences lead to disrupted operations, fish kills, and significant risks to human health. The conditions leading to blooms are driven by known, but complex processes. Heuristics exist about the drivers but the nonlinearity and opaqueness of relationships make it difficult to resolve using traditional rule-based mathematical models. An alternative approach leverages machine learning to uncover the conditions that lead to the closure of farms. This paper presents a comprehensive framework that combines semi-automated machine learning with ensemble classification approaches to predict site closures. Performance is evaluated on 7 years of site closure data from a shellfish farm in Southwest Portugal, together with publicly available environmental data. The model reports an accuracy of 83% across a challenging forecasting task. The proposed framework provides a pragmatic, scalable, site-specific decision tool to help aquaculture stakeholders mitigate the impacts of HABs.

Keywords Machine learning · Harmful algal blooms · Aquaculture · AutoAI

Handling Editor: Gavin Burnell

✉ Fearghal O'Donncha
feardonn@ie.ibm.com

Albert Akhriev
albert_akhriev@ie.ibm.com

Bruno Fragoso
fragoso.b@gmail.com

John Icely
john.icely@gmail.com

¹ IBM Research Europe, Dublin, Ireland

² Sagremarisco Lda, 8650-999 Vila do Bispo, Portugal

³ University of Algarve, 8005-139 Faro, Portugal

Introduction

Harmful algal blooms (HABs) are arguably the greatest threat to inland water quality, public health, and aquatic ecosystems (Finnis et al. 2017; Cruz et al. 2021). HABs are particularly damaging for the aquaculture industry causing extensive damage to operations. The effects on farmed finfish include physical interference, deoxygenation, or ichthyotoxicity (Davidson et al. 2021). Recently, large blooms of the common algae *Chrysochromulina leadbeateri* killed 8 million salmon in ocean net pens in Norway (Karlson et al. 2021). Shellfish farmers are potentially more exposed to HAB effects since toxins consumed by the filter feeders are bioaccumulated and bioamplified throughout the food web to affect humans and other organisms (Van Dolah 2000). There are several diseases associated with shellfish toxins including paralytic shellfish poisoning (PSP), diarrhetic shellfish poisoning (DSP), and amnesic shellfish poisoning (ASP) (Basti et al. 2018).

This significant risk to human health has led to a rigorous monitoring and management framework to minimize human health risk (Trainer and Hardy 2015; Trainer 2020; Harley et al. 2020; Legleiter et al. 2022). Monitoring of phytoplankton communities in the marine environment is required by the European Union (EU) legislation, such as the Marine Strategy Framework Directive (MSFD, 2008/56/EC and 2017/ 845/EC, 2008) and Water Framework Directive (WFD, 2000/60/EC), as well as the national law of the member states (Garmendia et al. 2013). If the concentration of shellfish biotoxins exceeds threshold levels, harvesting restrictions are applied until toxins are reduced (Davidson et al. 2021).

Monitoring and early warning of adverse harmful algal blooms is of critical importance for aquaculture industry stakeholders and consequently has been the focus of much research effort (O'Donncha and Grant 2019). These include comprehensive in-situ monitoring programs, the use of remote sensing products, and sophisticated computer modeling approaches that can be used to detect and forecast blooms (e.g., Wynne et al. (2020); Hardison et al. (2019)). Traditionally, computationally expensive, large-scale, physics-based models have been used to simulate algal concentrations (or a suitable proxy). The limitations to these include the high degree of user skill required and the difficulty to parametrize across different geographical regions (McGillicuddy 2010). Instead, the task of forecasting HAB events is ideally suited to a data-driven machine learning approach — especially since neither algae species data nor site-specific hydrodynamic/thermodynamic data are required.

This paper details a machine learning framework to forecast toxin events at a shellfish aquaculture farm site in Portugal. The objective is to develop an early warning system for farmers to forecast closures of shellfish sites due to elevated toxin levels in shellfish flesh. A forecasting model was trained using historic data from the Portuguese Institute of Sea and Atmosphere (IPMA) on-site closures (Portuguese Institute for Sea and Atmosphere 2022) together with environmental variables. Contributions of the paper are as follows:

- A transferable framework based on ocean data and AutoAI models to provide early warning of toxin events.
- A robust feature engineering approach that is amenable to the complexities of ocean time series datasets
- Evaluation of the approach on data from an operational shellfish site in South-West Portugal.

Related work

Early warning and accurate forecasting of algal blooms hold significant importance for public health organizations, fish farmers, tourist organizations, and various other stakeholders. As a result, substantial research efforts are dedicated to developing operational forecasting models capable of accurately capturing the intricacies of algal bloom dynamics.

Numerous algal forecasting models rely on numerically resolving the complex processes involved in the formation and development of algal blooms. This entails understanding the circulation patterns of the ocean and modeling the kinetics of individual algae species. Consequently, this scientific challenge is notoriously difficult (Roiha et al. 2010).

Some studies simplify the problem by focusing solely on circulation patterns, treating algae as passive particles transported by currents (Pinto et al. 2016). In contrast, more comprehensive approaches take into account both circulation and biology. One notable example is the Gulf of Mexico Harmful Algal Bloom Operational Forecast System (GOMX HAB-OFS), which provides a 10-day forecast of the “red tide” caused by *Karenia brevis*. This system combines a Regional Ocean Modeling System circulation model with a biological sub-model that considers cyst germination, cell growth rates, and other factors (Kavanaugh et al. 2013). By integrating satellite imagery, in-situ monitoring, and hydrodynamic modeling, the system estimates the current extent and intensity of the bloom, predicts its trajectory, and forecasts concentration levels.

A similar framework is employed to forecast cyanobacteria blooms in Lake Erie (Wynne et al. 2011). However, a primary challenge with these approaches is the computational expense involved in implementing them at high resolution across large spatial and temporal scales, often requiring high-performance computing facilities (O’Donncha et al. 2020). Additionally, configuring and parameterizing these models is a complex task typically requiring extensive expertise (O’Donncha et al. 2015).

To overcome these limitations, recent years have witnessed a growing interest in utilizing machine learning (ML) to develop cost-effective approximations or surrogates of physics-based models (Lary et al. 2004; Ashkezari et al. 2016; James et al. 2018; O’Donncha et al. 2019). Traditionally, data mining approaches have been employed to identify areas prone to algal blooms by determining influential features in bloom formation (Gokaraju et al. 2011; Chau and Muttill 2007). However, more recent attention has shifted towards ML methods that employ regression-based approaches to forecast future algal bloom concentrations.

Park et al. (2015) implemented an artificial neural network (ANN) and support vector machine (SVM) networks to predict Chl-*a* concentration in the Juam Reservoir and Yeongsan Reservoir. Using weekly measurements of water quality (Chl-*a*, phosphate phosphorus, ammonium nitrogen, nitrate nitrogen, and water temperature), and meteorological data (solar radiation and wind speed) over a 7-year period as input data, both the SVM and ANN made seven-day-ahead predictions of Chl-*a* concentrations. While these models indicated good predictive skill, they relied on difficult-to-collect water-quality observations that are not widely available.

Zhang et al. (2016) implemented a five-layered neural network to forecast algal blooms in the coastal waters of East China. The model was trained on 4 years of observed water-quality variables, including temperature, salinity, pH, Chl-*a*, chemical oxygen demand, dissolved oxygen, phosphate, acid nitrate, nitrate, ammonia nitrogen, and silicate, to forecast phytoplankton density as a proxy for algae formation.

In their study, Lee and Lee (2018) examined the performance of machine learning (multilayer perceptron, MLP) and deep learning models (recurrent neural network, RNN; and

long short-term memory, LSTM) to forecast Chl-*a* (as a proxy for algal blooms) in four rivers in South Korea. The models were applied to 16 monitoring stations, and the results revealed that ordinary least squares regression outperformed the deep learning models at five locations. Among the deep learning models, MLP and RNN achieved the lowest root mean square error (RMSE) at four and three locations, respectively, while LSTM performed best at four stations. However, a more recent study by Wolff et al. (2020) reported that simpler models such as generalized additive models (GAM) or random forest (RF) performed better than more sophisticated approaches like LSTM.

Fernandes-Salvador et al. (2021) conducted a comprehensive review study on harmful algal bloom (HAB) issues in European Atlantic waters and the status of early warning systems. They categorized the early warning systems into five major types: industry alert “bulletin” reports, particle tracking-based systems, statistical models based on remote sensing, statistical and machine learning models based on the fusion of multiple data sources, and mechanistic full-low trophic ecosystem models. These systems often combine multiple methods and are typically communicated to end users through expert interpretation.

CoastObs (2023) offers a HAB forecasting service that predicts the probability of a toxic bloom causing the closure of a production area. Their system employs a combination of advanced and basic models. However, they highlight the requirement of measuring nutrient concentrations, which is currently not available through existing monitoring programs or publicly accessible data. We believe that this paper is the first to predict the likelihood of farm closure based on publicly available data.

Methodology

Study site

The site for this study is located at the intersection of the west and south coast of Portugal near Sagres (37° 00' N, 8° 53' W). This coast has a narrow continental shelf that descends rapidly to depths of over 1000 m at the continental slope (Fig. 1). There are no perennial rivers, but the area is affected by coastal upwelling events which promote high primary production due to the upwelling conditions induced by northerly winds, occurring mostly during early spring to late summer. The cold, nutrient-rich water from these events stimulates high primary productivity that has enabled the development of offshore aquaculture for bivalves. More detailed information about this region can be found in the following articles and references therein: (Cravo et al. 2010; Krug et al. 2017; Danchenko et al. 2019; Santos et al. 2021; Danchenko et al. 2022; Icely and Fragoso 2023).

Since 1985, IPMA (the Portuguese Institute for the Sea and Atmosphere) has been responsible for the national monitoring program for HAB to identify potential levels of biotoxins in shellfish (Silva et al. 2016). With a funding contribution from the EU through the ASIMUTH project (Applied Simulations and Integrated Modeling for the Understanding of Toxic and Harmful Algal Blooms), the Portuguese have developed a weekly bulletin that provides information on the closure status of shellfish harvesting areas throughout Portugal (Silva et al. 2016; Fernandes-Salvador et al. 2021). The shellfish areas have been divided up into specific zones; in the case of the Sagres site, the zone is L7c (Figure 1 from Danchenko et al. (2019)). However, the L7c zone was split at the end of 2018 into L7c1 and L7c2; the Sagres site is within L7c1 (Portuguese Institute for Sea and Atmosphere 2022). Thus, for the analysis in

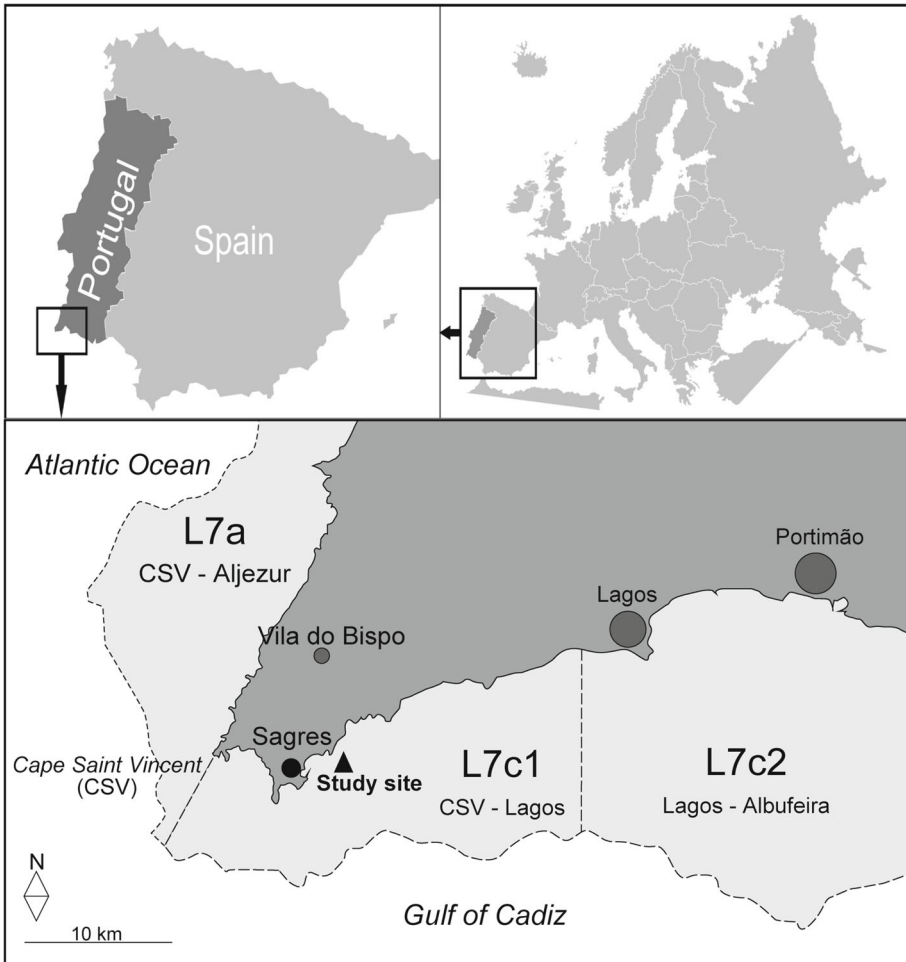


Fig. 1 Study site at the offshore mussel farm at Sagres SW Iberia (Portugal); distances from IPMA phytoplankton monitoring stations (figure from Danchenko et al. (2019))

this paper, closures over a period of 7 years were considered for zone L7c from August 2014 until 2018 and then subsequently for L7c1 up to May 2021.

Input data

The presence of toxins in shellfish flesh are regularly monitored by the Portuguese Institute of Sea and Atmosphere (IPMA). When concentrations exceed threshold levels, the shellfish farm is closed until values return to acceptable limits. For each of PST, AST, and DST, the legal thresholds are 800µg, 20 mg, and 160µg equivalent per kg shellfish, respectively (Portuguese Institute for Sea and Atmosphere 2023).

Data on licensed area closures are provided by IPMA from August 2014 and provide daily information on whether a site is open or closed.

Table 1 Summary of the data sources used in this study and their respective resolution

Data	Horizontal resolution	Temporal resolution	Source
Weather data	4 km	Hourly	(Villali 2021)
Copernicus ocean data	1.8 km	Hourly	(Tonani et al. 2019)
Site closures	Bay-scale	Daily	(Silva et al. 2016)

Fig. 2 provides a full list of variables together with statistical relationships, while Fig. 3 summarizes the data variation

The environmental data used for this study consisted of ocean product data from the Copernicus Marine Service model repository (Tonani et al. 2019), and weather data from IBM Environmental Intelligence Suite (Villali 2021). While data were collected or generated at different resolutions, we resampled all variables to daily intervals to correspond to site closure values. Table 1 details the data sources and resolutions used for this study, while Figs. 2 and 3 provides a summary of the variables.

The data utilized in this research is publicly accessible. Although we are not authorized to distribute weather data, you can acquire a free API key from the supplier to retrieve the data. Following a complimentary registration, Copernicus ocean product data can be obtained from the marine services portal. Additionally, the full source code for downloading the data,

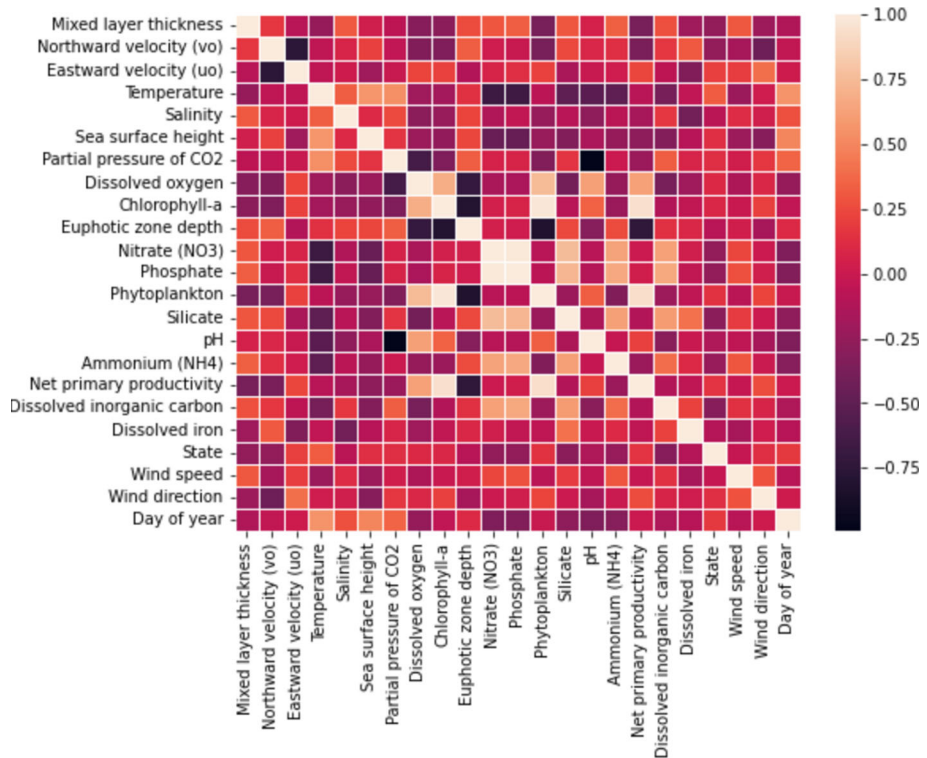


Fig. 2 Correlation matrix of all variables considered for this study. The name of each variable is denoted on the axes, while the correlation between each variable pair is provided in the matrix plot

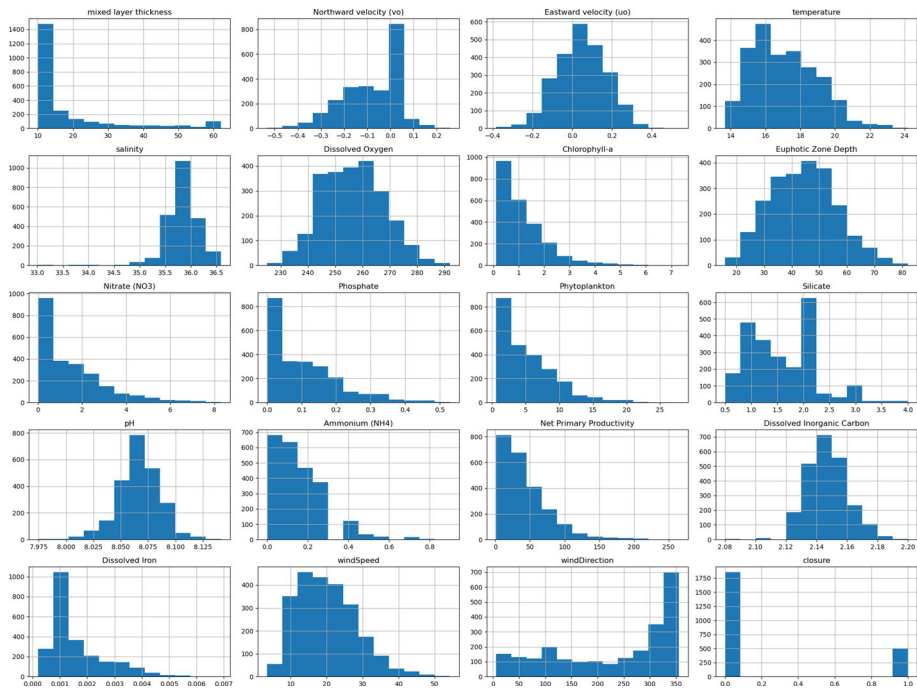


Fig. 3 Plot of the data distribution for all variables. The *state* variable indicates whether the aquaculture site is closed (1) due to elevated levels of biotoxin

implementing the machine learning models, and executing the analysis has been publicly released at: https://github.com/fearghalodomncha/habs_ml/.

Machine learning

Classical works in machine learning and optimization, introduced the “no free lunch” theorem (Wolpert and Macready 1997), demonstrating that no single machine learning algorithm can be universally better than any other in all domains — in effect, one must try multiple models and find one that works best for a particular problem. Selection of the most suitable algorithm and algorithmic settings is one of the most complex aspects of machine learning applications and is highly dependent on user skill. A powerful approach to select optimal algorithms is *automatic machine learning* (AutoML) frameworks that aim to *learn how to learn* (Drori et al. 2018). AutoML tools use a variety of techniques, such as differentiable programming, tree search, evolutionary algorithms, and Bayesian optimization, to find the best machine learning pipelines for a given task and dataset (Drori et al. 2018). We used the open-source Lale Python library (Hirzel et al. 2019) for automated machine learning to simplify model development. Lale is designed to automatically select algorithms, tune hyperparameters, and explore pipeline topologies from a set of available preprocessors and machine learning algorithms suggested by the user.

The algorithms considered for this study were Random Forest (RF), Gradient Boosting (GradBoosts), XGBoost (XGB), and MultiLayer Perceptron (MLP). The first three models are from the classification and regression tree (CART) family, based on the *aggregation*

of a large number of decision trees. Decision trees are a conceptually simple yet powerful prediction tool that breaks down a dataset into smaller and smaller subsets while at the same time, an associated decision tree is incrementally developed. The resulting intuitive pathway from explanatory variables to outcome serves to provide an easily interpretable model. The combination of simplicity, robustness, and interpretability makes these an extremely popular tool in machine learning.

In RF (Breiman 2001), each tree is a standard Classification or Regression Tree (CART) that uses what is termed node “impurity” as a splitting criterion and selects the splitting predictor from a randomly selected subset of predictors (the subset is different at each split). Each node in the regression tree corresponds to the average of the response within the sub-domains of the features corresponding to that node. The node impurity gives a measure of how badly the observations at a given node fit the model. In regression trees, this is typically measured by the residual sum of squares within that node. Each tree is constructed from a bootstrap sample drawn with replacement from the original data set, and the predictions of all trees are finally aggregated through majority voting (Boulesteix et al. 2012). RF is especially popular for its strong performance with little hyperparameter tuning (i.e., works well with the default values specified in the software library).

GradBoost and XGBoost primarily differ from RF in how decision trees are built. While they both share many characteristics and advantages with RF (namely interpretability, predictive performance, and simplicity), a key difference facilitating performance gain in boosting methods is that decision trees are built *sequentially* rather than *independently*. This allows each new tree help to correct errors made by previously trained trees (Breiman 1997).

The XGBoost algorithm was developed at the University of Washington in 2016 and since its introduction has been credited with winning numerous Kaggle competitions and being used in multiple industry applications. XGBoost provides algorithmic improvements such as a sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning, together with optimization towards distributed computing, to build a scalable tree boosting system that can process billions of examples (Chen and Guestrin 2016).

The multi-layer perceptron (MLP) model on the other hand comes from the neural network family and is loosely based on the anatomy of the brain. Such an artificial neural network is composed of densely interconnected information-processing nodes organized into layers. The connections between nodes are assigned “weights,” which determine how much a given node’s output will contribute to the next node’s computation. During training, where the network is presented with examples of the computation it is learning to perform (i.e., likelihood of HAB event), those weights are optimized until the output of the network’s last layer consistently approximates the training data set (i.e., correctly predicts a shellfish site closure) (Hornik et al. 1989). A significant drawback of neural network methods is that the explainability of predictions is challenging, particularly as the size of the network increases.

Model setup and training

Since in-situ and remote sensing data did not provide contiguous data over the entirety of the study period, we used ocean model data from the Copernicus Marine Service repository (Tonani et al. 2019) and weather data from the IBM Environmental Intelligence Suite historical reanalysis product (Villali 2021). Leveraging publicly available data as features also serves to enhance the generalisability of the approach. The framework presented here can easily be applied to any other region in Europe (or globally where high-resolution ocean data is available). Data were resampled to daily values and combined with our *label* data inform-

ing on whether the shellfish site was open (denoted zero) or closed (denoted one). Label data from IPMA were available since August 2014, hence our study covered the period 05/08/2014–01/05/2021. The day-of-year was included as a feature to represent temporal variations in regional HAB developments. In order to augment the model's ability to assimilate temporal dynamics, we encoded these features through the application of trigonometric functions, thereby effectively encapsulating their inherent cyclical nature. The final design matrices were of dimension [2492, 20] features and a corresponding vector of length [2492] label data.

Figure 2 displays the correlation matrix for the data. The label data reporting when the shellfish site is designated open or closed is denoted by the *state* variable and the correlation against the 22 selected explanatory variables are presented. Naturally, the degree of correlation between constituents varies. The highest correlation is reported for physical variables such as temperature and velocity, biogeochemical variables such as net primary production, or temporal variables such as day-of-year. Figure 2 informs that correlation exists between some of these variables and it can be considered the task of the machine learning model to extract the characteristics of the relationships and use this information to craft a predictive model.

Data were split into 80% training and validation dataset and 20% testing. To avoid data leakage, the testing dataset considered a contiguous period from 28/08/2017 to 01/01/2019 with the remainder retained for training. Data leakage in timeseries forecasting can occur when temporal dependencies are not respected during the train-test split. If the data is shuffled or the test set includes data that comes before the training set, it can lead to leakage and unrealistic performance estimates.

Further, the selected testing period contained a relatively high number of closures which allowed a more comprehensive evaluation.

Figure 3 indicates that the shellfish site was closed 497 of 2353 days (21%). Post-splitting, the training dataset contained 326 closures from 1883 days (17%), while the test set contained 171 from 470 (36%). Imbalanced data can lead to model performance being skewed towards the majority class (by simply learning to output the majority class). Our selected split penalizes such a tendency by ensuring the test dataset is not dominated by the majority class. Importantly by focusing on selecting a “difficult” test dataset, we avoid potentially biasing model results.

To address imbalances in the training data, we instead adopted a statistical approach, termed data sampling. Fundamentally, this involves generating new data points for the minority class (up-sampling), and/or removing data points from the majority class (down-sampling). Implementing data sampling requires consideration of different combinations of up-sampling and down-sampling to improve model accuracy. Up-sampling proceeded by sampling points from the minority class with replacement (i.e., making copies of points from the minority class), while down-sampling involved deleting points from the majority class (He and Ma 2013). Adjusting the class balance does not introduce new data but simply adjusts the ratio either by replicating the minority class or by throwing away some of the majority class. Importantly, the sampling techniques were applied only to the training data and the test data were untouched to provide a comprehensive evaluation. The optimal resampling rate is a hyperparameter to be selected as part of model training.

Re-sampling provides a data-centric approach to address model inaccuracy. An alternative *model-centric* approach is based on the classic paradigm of combining multiple model predictions into a single classifier.

Ensemble learning has a relatively simple concept that aims to reduce the model error by combining multiple “weak” learners. By combining different, diverse models, the expected

error of the group reduces, while each individual model remains unchanged. Adding more models to the ensemble improves accuracy, under the condition that each learner is uncorrelated to the others. Of course, this is increasingly difficult to achieve and it is subject to the law of diminishing returns (Kyriakides and Margaritis 2019).

Two main categories of ensemble classification exist voting and stacking or blending. Voting, as the name implies, refers to techniques that allow models to vote in order to produce a single answer. The most popular (most voted for) answer is selected as the winner. The results of the voting can be decided based on the number of votes (hard mode) or on both the number of votes and probability returned by each ML model (soft mode). Stacking, on the other hand, refers to methods that utilize a model (the meta-learner) to learn how to best combine the base learner's predictions. Although stacking entails the generation of a new model, it does not affect the base learners, and instead, the new model aims to learn which combination of the base forecasts provides the best estimate.

The feature transformations considered were standardization and principle component analysis (PCA). Standardization is a scaling technique where the values are centered around the mean with a unit standard deviation. PCA is a data-driven modeling technique that transforms a set of correlated variables into a smaller set of uncorrelated variables while retaining most of the original information.

Model testing and evaluation

Model testing considered the ability of the model to accurately forecast the likelihood of site closure at the shellfish site due to an algal bloom event. A number of standard performance metrics are commonly used in classification models. True positives (TP) and true negatives (TN) report events that the model correctly predicts. Conversely, false positives (FP) and false negatives (FN) are events the model misdiagnoses. Oftentimes, an FN is the most damaging since the site suffers an unexpected toxin event.

Common classification model skill metrics include *accuracy*, *precision*, and *recall*. Accuracy (ACC) is simply defined as the ratio between the number of correct predictions divided by total number of predictions ($ACC = \frac{TP+TN}{N}$); precision or positive predictive value (PPV) is defined as the fraction of correct times the model returns positive ($PPV = \frac{TP}{TP+FP}$); recall or sensitivity quantifies the amount of times that the sites is actually closed that the model correctly predicts ($Recall = \frac{TP}{TP+FN}$).

Precision and recall are often in conflict and one must consider both metrics and how they contribute to the forecasting requirements. Recall provides a measure of the number of events that are correctly diagnosed, while precision measures the proportion of events flagged by the model that were correctly classified.

The F1 score is an evaluation metric that combines precision and recall and can be expressed as:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (1)$$

The F1 score is less influenced by TN than accuracy. In many applications, TN does not have significant business implications, whereas FN and FP often have operational implications.

Model evaluation adopted a 10-fold cross-validation. Cross-validation is a technique in which the model is trained using a subset of the data set and then evaluated using a complementary subset of the data set.

Results

The model was trained using the Lale semi-automated machine learning library (Hirzel et al. 2019). In short, a number of algorithms were interrogated by the Lale library which simplifies model setup for data scientists by searching over possible choices for hyperparameters, algorithms, and data standardization schemes. The models considered were RF, GradBoost, XGB, and MLP, which are described in detail in the “Machine learning” section. Figure 4a summarizes the feature transformations and algorithms provided as options to the learning algorithm, while Table 2 details the hyperparameter ranges provided as inputs to the Lale search routines.

The experimental approach involved investigating various sampling techniques to address data imbalance. A grid search strategy was employed, exploring different values of up-sampling and down-sampling within the range of 0 to 1, with an incremental step size of 0.05. The imblearn library was utilized for data resampling. After evaluating different options, the optimal model performance was achieved with an up-sampling rate of 0.5 and a down-sampling rate of 0.6. This means that the minority class was up-sampled until its number

Table 2 Hyperparameters and their ranges used for model design

Model	Hyperparameters
RF	# of estimators $\in [50, 1000]$ min samples split $\in [0, 1]$ max features $\in \{“auto,” “sqrt,” “log2,” \text{ and } “None”\}$ max depth $\in [1, 20]$
GradBoost	# of estimators $\in [50, 1000]$ min samples split $\in [0, 1]$ max features $\in [0, 1]$ learning rate $\in [0, 1]$ max depth $\in [1, 20]$
XGBoost	# of estimators $\in [50, 1000]$ learning rate $\in [0, 1]$ gamma $\in [0, 1]$ subsample $\in [0, 1]$ lambda $\in [0.001, 1]$ n_estimators $\in [0, 1000]$ max depth $\in [1, 20]$
MLP	# of layers $\in [1, 50]$ # of nodes/layer $\in [1, 1000]$ $\lambda \in [0.001, 1]$

See the “Machine learning” section for further details on each model details. Note that we used the Lale autoML library to implement the hyperparameter optimization schemes and selected the default range of values for each model. This tends to hide much of the complexity of hyperparameter optimization from the user. For additional details on implementing the models, please refer to our GitHub repository at https://github.com/fearghalodonncha/habs_ml

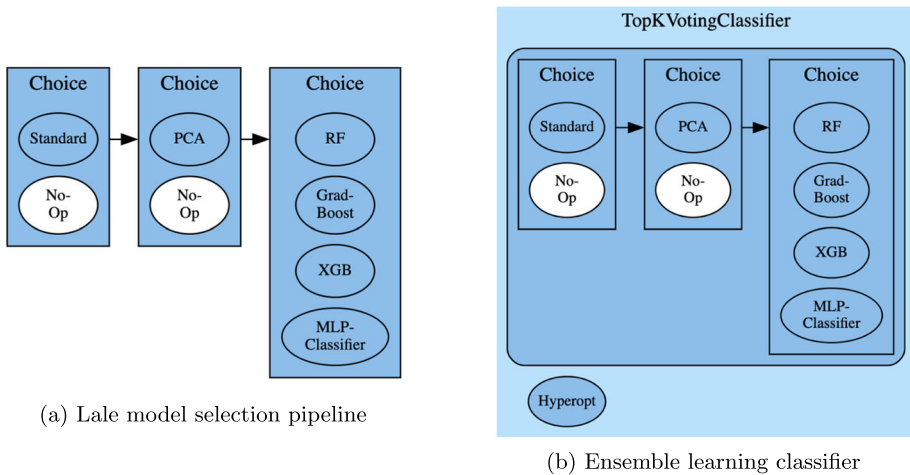


Fig. 4 Using Lale, we pass a variety of options related to feature transformation and algorithmic selection to the meta model. **a** presents the pipeline to select the best-performing individual model, while **b** presents the ensemble learning preprocessing and training process. The first box explores the effects of scaling the data to unit standard deviation; the second box considers whether PCA provides valuable transformations to the data; finally, the third box considers four different machine learning algorithms to identify the most appropriate algorithm or combination of such. Lale selects the pipeline that optimizes model skill using 10-fold cross validations

of instances reached 50% of the majority class, and then the majority class was reduced to achieve a ratio increase to 60%. Consequently, the ratio between the majority and minority classes in the training data transformed from 1557:326 prior to resampling to 1296:778 after the processing.

An ensemble model approach was implemented based on a voting approach that considered the same algorithms (RF, GradBoost, XGB, and MLP). Figure 4b summarizes the architecture of the meta-learner consisting of two different preprocessing options and the above four learning algorithms. Again Lale implemented the voting classifier algorithmic search to learn the optimal pipeline. From the set of all possible pipelines, Lale meta-learner computes the top-performing learners and implements an ensemble voting predictor.

Listing 1 details the optimal hyperparameter and pipeline topologies identified by Lale. For each individual model in the ensemble, it identified the optimal algorithm, preprocessing, and hyperparameter settings. Each of the 10 best-performing models is then combined using a voting approach. For the ensemble classifier, a hard or soft voting can be used with soft voting being selected in this study.

The ensemble learner was evaluated against two baseline models to benchmark model skill:

- Best-performing individual model trained on the original dataset with no resampling (called BPIM)
- Best-performing individual model trained on the resampled dataset that amended the ratio between majority and minority classes from 1557:326 to 1296:778 (called BPIM_S).

Listing 1 The optimal hyperparameter values, algorithm selection, and pipeline topologies identified. Hyperparameter and pipeline settings were automatically explored by the Lale library across 10000 iterations. We configured the model to select the 10 estimators or model pipelines that maximised performance. Models were combined using a voting method. Since only four algorithms (RF, GradBoost, XGBoost, and MLP) were provided to the model this meant that some of the algorithms contributed more than once to the prediction). A 'soft' voting method was selected by Lale which means that each vote is calculated based on the predicted probability of the output class.

```
# Print optimal settings selected for the Voting Ensemble Classifier
trained_ensemble.get_pipeline().pretty_print(ipython_display= True)

# Voting Ensemble optimal configurations
standard_0 = Standard(with_mean=False, with_std=False)
xgb = XGB(
    gamma=0.0010058465848622133,
    learning_rate=0.053017720054907694,
    max_depth=5,
    min_child_weight=2,
    n_estimators=546,
    reg_alpha=0.00021452714423669805,
    reg_lambda=0.18357919522200414,
    subsample=0.903408656659564,
)
standard_1 = Standard(with_mean=False, with_std=False)
rf_1 = RF(
    max_depth=3,
    max_features="log2",
    min_samples_leaf=0.10209804132000078,
    min_samples_split=0.22233307291641027,
    n_estimators=19,
)
... # additional models not presented for brevity
pipeline = VotingClassifier(
    estimators=[
        ("p2", Standard() >> pca >> xgb),
        ("p7", standard_0 >> NoOp() >> rf_1),
        ("p1", standard_3 >> NoOp() >> mlp_classifier),
        ("p9", NoOp() >> NoOp() >> rf_2),
        ..... # additional pipelines not shown
    ],
    voting="soft",
)
```

The two baseline models were compared against the final ensemble model trained on resampled data (called Ensemble). The model evaluation focused on predictive skill and robustness. Tables 3 and 4 summarize the relative performance of the different model implementations. Table 3 presents a confusion matrix for the three implementations. A confusion matrix is a common method for describing the performance of a classification model. This is a simple cross-tabulation of the observed and predicted classes for the data. The main diagonal denotes cases where the classes are correctly predicted (TP and TN) while the

Table 3 Confusion matrix reporting the predictive skill of the models

		Predicted condition					
		BPIM		BPIM_S		Ensemble	
		Closed	Open	Closed	Open	Closed	Open
True Condition	Closed	39	132	75	96	100	71
	Open	11	288	4	295	11	288

We present results for the best-performing model on the original data (BPIM), the best-performing model on resampled data to reduce biases (BPIM_S), and the ensemble classification model trained on the resampled data (Ensemble). For each model, the main diagonal reports the amount of times that the model correctly predicts when a site is actually closed, while the counter diagonal indicates incorrect prediction

counter diagonal illustrates the number of errors for each possible case (FP and FN). Results demonstrate that model BPIM provided a relatively accurate prediction of the site being open (correct prediction 288 out of 299 times), but failed to diagnose closure events. In short, the imbalanced nature of the data led the model to be biased towards positive events.

Table 4 summarizes the accuracy metrics reported by the three models. Accuracy ranges from 0.7 to 0.83 with the ensemble model significantly outperforming BPIM. The ensemble model reports a moderate drop in precision compared to BPIM_S. Namely, while the number of TP increases and FN decreases, the number of TN increases which leads to a moderate drop in precision. However, this reduction in FN is critical to practical model performance. Operators are typically accepting of a slight increase in the number of false notifications of a potential toxin event if the model accurately forecasts closures. Typically, the more severe the repercussions of an event are, the more important recall becomes as a metric.

A fundamental characteristic of data-driven approaches is that the model learns the patterns that enable it to make a prediction: instead of the expert encoding these relationships using physical equations, the model learns the most appropriate mapping between the given inputs or features and outputs or labels. Aquaculture industry stakeholders have been particularly interested to know what specific drivers contributed to predicting potential HAB events. Interpretable or explainable AI (Samek et al. 2019) are emerging topics in data science that aim to guide AI model interrogation.

Figure 5 presents the *feature importance* of the supplied data to the response variable or model prediction at the study site. Feature importance was computed based on the Shapley value (Cohen et al. 2005). The Shapley value derives from game theory Shapley (1997) and

Table 4 Accuracy and model skill reported for our three different model implementations

	Accuracy $\frac{TP+TN}{N}$	Precision (P) $\frac{TP}{(TP+FP)}$	Recall (R) $\frac{TP}{(TP+FN)}$	F1 score $\frac{2x(P*R)}{(P+R)}$
BPIM	0.7	0.78	0.23	0.35
BPIM_S	0.79	0.95	0.44	0.6
Ensemble	0.83	0.9	0.58	0.75

We compare two variants of the best-performing individual model (BPIM) against an ensemble classification approach. Accuracy is simply defined as the ratio between the number of correct predictions and total number of predictions. Precision or positive predictive value (PPV) is defined as the fraction of correct times the model returns positive. Recall or sensitivity quantifies the amount of times that the sites is actually closed that the model correctly predicts

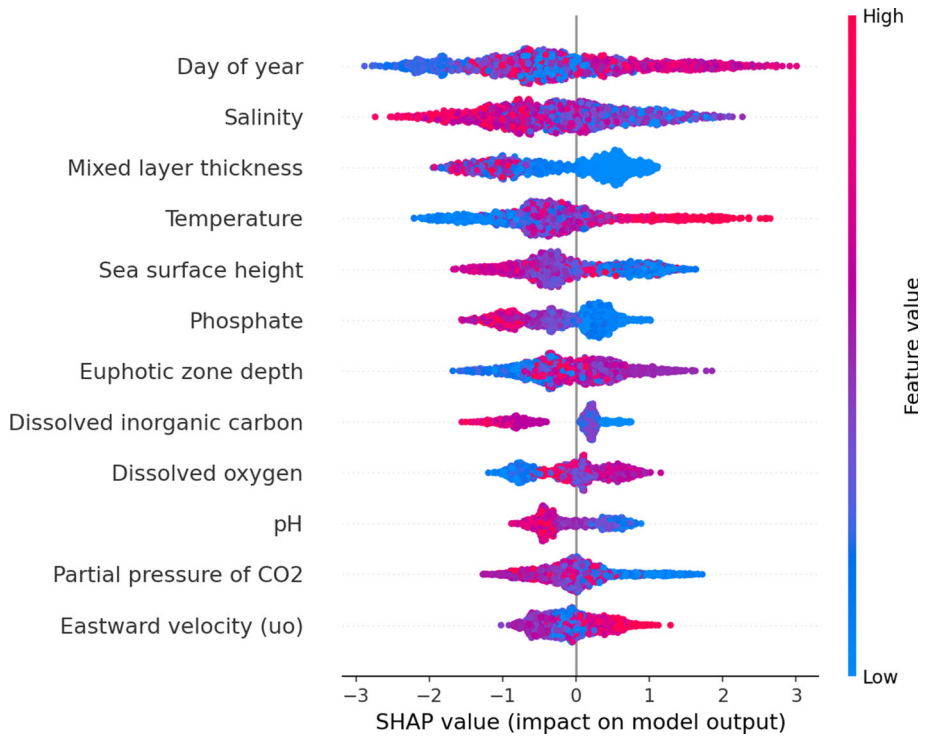


Fig. 5 Feature importance reported for the XGB model when predicting the likelihood that the shellfish site will be closed or not due to a toxin event. Visualization is limited to the 12 most important features. The y-axis reports the ranked list of features that contributed the most to variation prediction (site closure), while the x-axis presents the relative magnitude of those contributions. Ranking predictors in this manner can quickly help sift through large datasets and understand data trends (Kuhn et al. 2013)

aims to assign payout (or importance) to players (or features) depending on their contribution to the total payout. The feature importance measure computes the contribution or importance of each feature by calculating the increase of the model’s prediction error after permuting the feature. A feature is “important” if permuting its values increases the model error because the model relied on the feature for the prediction. A feature is “unimportant” if permuting its values keeps the model error unchanged because the model ignored the feature for the prediction (Breiman 2001).

Given the difficulties associated with extracting primary features from ensemble methods, we opted for BPIM_S in this analysis. Specifically, we employed the best-performing XGBoost model, which is highly suitable for explainability analysis. This choice underscores the significance of selecting an appropriate model that aligns with the specific requirements of a study, where factors like accuracy, explainability, computational cost, and robustness often play crucial roles.

Figure 5 illustrates that the day-of-year is the most important contributor to prediction. This is not surprising, as many papers on this site (Loureiro et al. 2005; Goela et al. 2014; Cristina et al. 2016; Silva et al. 2016; Danchenko et al. 2019, 2022; Fernandes-Salvador et al. 2021), show that the phytoplankton community with HAB species changes from diatoms in early spring to summer, during the upwelling season, to dinoflagellates during the relaxation

of upwelling in autumn. The other primary contributory features include physical variables such as temperature and salinity, as well as important biogeochemical variables, including phosphate, dissolved oxygen, and the Net Primary Productivity of Carbon. Notably eastward (but not northward) velocity is also an important contributor to prediction. HAB events in the region are strongly influenced by upwelling events (Garmendia et al. 2013) driven by flows from offshore. Elucidating the complexity of prediction, these upwelling events are driven by physical processes such as wind speed and ocean flows which significantly modify properties such as nutrient levels, dissolved oxygen, and temperature. Interpretability approaches provide a framework to disentangle these complex events to some degree, but they must ultimately be informed by oceanography, biology, and domain expertise.

Discussion

These results illustrate how machine learning can improve decision-making on shellfish farms. Experienced farm operators have heuristic knowledge of the environmental conditions that lead to HAB events (upwelling processes, ambient wind conditions, water temperature range, etc.). Machine learning offers the opportunity to mine these processes and implement automated frameworks to identify high-risk periods for HAB-induced closures.

Relying only on public datasets, the model can be seamlessly transported to other locations. Naturally, predictive skills can likely be improved by incorporating long-term in-situ monitoring data that more closely captures system dynamics. However, increasing the volume of data is often the optimal strategy to improve model performance. This alludes to another advantage of the framework — the model will continue to improve as more data is collected.

The paper details the key considerations to apply machine learning to complex environmental datasets. Many aspects of the proposed framework are relevant for other environmental studies, namely:

- The design of suitable features to represent the dynamics of the system is critical. This includes the choice of suitable variables to represent important processes — often this is driven by domain expertise; understanding the statistical relationships between variables that can be used to inform feature selection; and implementing robust feature engineering and transformation routines to avoid data scale inconsistencies.
- The “no free lunch” theorem in classical optimization (Wolpert and Macready 1997), indicates that no single optimization algorithm is superior to all others. For practical purposes, one must explore different algorithms and decide on the one that works best for the problem at hand. While algorithmic search can be done manually by the data scientist, AutoAI approaches are a valuable tool to do this easily, at scale.
- Explainable and interpretable AI methods are extremely valuable for machine learning-based analysis of environmental processes. It is vital that the relationships used by the model to make a prediction agree with those expected by the domain experts. Otherwise, the grounding of the model is flawed and will not perform well in unseen situations.
- Many environmental studies deal with the prediction or evaluation of extreme events. However, machine learning is predicated on two core assumptions: 1) maximizing accuracy is the goal, and 2) in use, the classifier will operate on data drawn from the same distribution as the training data. Consequently, careful attention is required when the prediction classes are highly imbalanced. Naïve classifiers which always predict the majority class can achieve high accuracy in this case with no model skill.

- Artificially rebalancing the data by up- or down-scaling is a robust technique to improve prediction (Provost 2000). Results presented in this paper demonstrate that it produces significant model uplift for predicting adverse environmental conditions. Further, the ease of implementation makes it an important part of the data scientist's toolbox.
- Ensemble prediction is a widely used technique to improve model performance by considering forecasts from multiple different models. The fundamental objective of ensemble forecasting is to investigate inherent uncertainty to provide more accurate information about future states. Classical works on ensemble forecasting demonstrated that the ensemble mean should give a better forecast than a single deterministic forecast as long as the ensemble represents the uncertainty present in the forecast (Epstein 1969; Leith 1974). Ensemble methods with perturbed initial conditions are ubiquitous in meteorology and these focus on quantifying the fastest-growing errors with techniques such as the breeder method and optimal perturbation analysis, common in data-assimilation implementations (Turner et al. 2008). Multi-algorithm ensembles in machine learning provide a robust framework to explore model uncertainty and serve as a means of regularization to the model.

This paper illustrates the potential value of machine learning as a decision-support tool for aquaculture. A holistic experimental framework is vital to achieve a high-performing model. Table 4 summarizes the performance of three different models. While accuracy is high for all three models, this is only a useful metric when there is an equal distribution of classes. In many applications (including this), recall and precision scores provide a more representative measure. Recall measures the number of actual closures that were accurately predicted while the F1 score is defined as the harmonic mean of precision and recall.

In this study, intelligent feature engineering and algorithmic selection increased recall from 0.23 to 0.58 while the F1 score increased from 0.35 to 0.75. This means that almost 60% of site closures were correctly diagnosed. Importantly, the final model reported a precision of 0.9 indicating that the model reported a relatively small number of false positives (reporting site closure when in fact it was open). Considering the complexities of the problem this is an excellent model performance. It can drastically improve the ability of operators to respond to adverse events. Currently, farmers have no real insight into the likelihood of closures beyond heuristics and experience (O'Donncha and Grant 2019). With a robust early forecasting system, operators can make decisions to ameliorate the effects. Potential decisions include: harvest the molluscs early to avoid the toxin event, move the installation to another location that is less exposed, or incorporate the upcoming disruption into their scheduling and amend harvesting and product sales timelines.

This paper considers the problem as a timeseries case. Of course, processes are also influenced by conditions at adjacent locations. Further research could explore the integration of spatial dependencies. Examples include the use of LSTM networks to represent spatial relationships (O'Donncha et al. 2022), while convolutional neural networks (CNN) are a powerful deep learning tool that demonstrates exceptional performance processing image data. Combining CNN with time series machine learning models has previously been used to forecast ocean temperature Yang et al. (2017). This can allow the model to more effectively learn spatial and temporal drivers of HAB developments.

In the future, further investigations in this field will aim to incorporate a more explicit depiction of the spatial connections among blooms. Graph neural networks (GNN) offer a potent technique that permits the integration of heuristic or statistical details about physical properties and the organization of the structure within the modeling framework (Langbridge et al. 2023). The benefit of this approach is that heuristic relationships or modeled flow

patterns can be utilized to develop a graph topology that links and informs about harmful algal bloom (HAB) incidents in various farming locations.

It is worth noting that the drivers behind HAB events are likely to exhibit similarities across different sites. This presents a clear opportunity to enhance the performance and generalization capabilities of machine learning models by leveraging transfer learning techniques (Oruche and O'Donncha 2023). In particular, transformer architectures can be trained on large datasets to learn spatial and temporal dependencies in HAB occurrences. Although initially designed for natural language processing (NLP) tasks, transformers have been successfully applied to other domains, such as computer vision and graph modeling, due to their powerful capabilities.

One advantage of using transformers for this task is their ability to handle large amounts of data and complex relationships between variables. Transformers can learn patterns across both time and space, allowing them to model the complex interactions between different environmental factors that contribute to algal blooms. Additionally, transformers can learn from multiple modalities of data, such as satellite images and oceanographic data, and combine them into a unified model.

Another advantage of using transformers for this task is their ability to scale to large datasets. As the amount of data related to algal blooms continues to grow, transformers can be trained on increasingly larger datasets without sacrificing performance.

Conclusions

This paper introduces a framework designed to forecast the closure of shellfish sites caused by excessive levels of harmful toxins. The framework encompasses data conditioning, the development of machine learning (ML) models, and ensemble model forecasting, while also addressing pragmatic aspects of model development and implementation.

The presented model demonstrates high predictive accuracy when appropriate data conditioning and algorithmic selection are applied. This research holds significant practical value for shellfish operations, as timely information is crucial for effective decision-making. By predicting site closures based on environmental conditions, the framework aligns its forecasts with available environmental predictions.

The approach relies on publicly available data and leverages robust autoML metalearners to automate algorithm selection and parameterization. This feature enables easy deployment of the framework to other shellfish sites. Moreover, the computational efficiency of the approach makes it suitable for on-site or edge deployment.

Presentations of this study to individual stakeholders and the Portuguese Association of Aquaculture Farmers have received highly positive feedback. In fact, some stakeholders have expressed interest in expanding the approach further. They have inquired about the potential for predicting bivalve larvae spawning and settlement, as well as forecasting the optimal conditions for achieving maximum sales returns of the product.

Author contribution F.O'D., B.F., and J.I. wrote the main manuscript text and F.O'D. prepared figures. A.A. implemented scientific analysis. All authors reviewed the manuscript.

Funding Elements of this research received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 773330.

Data availability All data and code used for this study can be downloaded from https://github.com/fearghalodonncha/habs_ml.

Declarations

Conflict of interest The authors declare no competing interests.

References

- Ashkezari MD, Hill CN, Follett CN et al (2016) Oceanic eddy detection and lifetime forecast using machine learning methods. *Geophys Res Lett* 43(23):12–234. <https://doi.org/10.1002/2016GL071269>
- Basti L, Hégaret H, Shumway SE (2018) Harmful algal blooms and shellfish. In: Shumway SE, Burkholder JM, Morton SL (eds) *Harmful algal blooms: a compendium desk reference*. John Wiley & Sons, Hoboken, NJ, USA, pp 135–190. <https://doi.org/10.1002/9781118994672.ch4>
- Boulesteix AL, Janitza S, Kruppa J et al (2012) Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *WIREs Data Mining Knowl Discov* 2(6):493–507. <https://doi.org/10.1002/widm.1072>
- Breiman L (1997) Arcing the edge. Tech. rep., University of California, Berkeley. <https://statistics.berkeley.edu/tech-reports/486>
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32. <https://doi.org/10.1023/A:1010933404324>
- Chau KW, Muttill N (2007) Data mining and multivariate statistical analysis for ecological system in coastal waters. *J Hydroinf* 9(4):305–317. <https://doi.org/10.2166/hydro.2007.003>
- Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proc of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, New York, USA, pp 785–794. <https://doi.org/10.1145/2939672.2939785>
- CoastObs (2023) CoastObs - real services for real people. <https://coastobs.eu/services/harmful-algal-blooms-hab-forecasts>
- Cohen S, Ruppim E, Dror G (2005) Feature selection based on the shapley value. International joint conference on artificial intelligence, pp 1–6. <https://www.ijcai.org/Proceedings/05/Papers/0763.pdf>
- Cravo A, Relvas P, Cardeira S et al (2010) An upwelling filament off southwest Iberia: effect on the chlorophyll a and nutrient export. *Cont Shelf Res* 30(15):1601–1613. <https://doi.org/10.1016/j.csr.2010.06.007>
- Cristina S, Cordeiro C, Lavender S et al (2016) MERIS phytoplankton time series products from the SW Iberian Peninsula (Sagres) using seasonal-trend decomposition based on Loess. *Remote Sens* 8(6):449. <https://doi.org/10.3390/rs8060449>
- Cruz RC, Reis Costa P, Vinga S et al (2021) A review of recent machine learning advances for forecasting harmful algal blooms and shellfish contamination. *J Mar Sci Eng* 9(3):283. <https://doi.org/10.3390/jmse9030283>
- Danchenko S, Fragoso B, Guillebault D et al (2019) Harmful phytoplankton diversity and dynamics in an upwelling region (Sagres, SW Portugal) revealed by ribosomal RNA microarray combined with microscopy. *Harmful Algae* 82:52–71. <https://doi.org/10.1016/j.hal.2018.12.002>
- Danchenko S, Dodge JD, Icely JD et al (2022) Dinoflagellate assemblages in the West Iberian upwelling region (Sagres, Portugal) during 1994–2001. *Front Mar Sci* 9:591759. <https://doi.org/10.3389/fmars.2022.591759>
- Davidson K, Whyte C, Aleynik D et al (2021) HABreports: online early warning of harmful algal and biotoxin risk for the Scottish shellfish and finfish aquaculture industries. *Front Mar Sci* 8:350. <https://doi.org/10.3389/fmars.2021.631732>
- Drori I, Krishnamurthy Y, Rampin R et al (2018) AlphaD3M: machine learning pipeline synthesis. In: *AutoML Workshop at ICML*, pp 1–8. <https://doi.org/10.48550/arXiv.2111.02508>
- Epstein ES (1969) Stochastic dynamic prediction. *Tellus* 21(6):739–759. <https://doi.org/10.1111/j.2153-3490.1969.tb00483.x>
- Fernandes-Salvador JA, Davidson K, Sourisseau M et al (2021) Current status of forecasting toxic harmful algae for the North-East Atlantic shellfish aquaculture industry. *Front Mar Sci* 8:656. <https://doi.org/10.3389/fmars.2021.666583>
- Finnis S, Krstic N, McIntyre L et al (2017) Spatiotemporal patterns of paralytic shellfish toxins and their relationships with environmental variables in British Columbia, Canada from 2002 to 2012. *Environ Res* 156:190–200. <https://doi.org/10.1016/j.envres.2017.03.012>
- Garmendia M, Borja Á, Franco J et al (2013) Phytoplankton composition indicators for the assessment of eutrophication in marine waters: present state and challenges within the European directives. *Mar Pollut Bull* 66(1–2):7–16. <https://doi.org/10.1016/j.marpolbul.2012.10.005>
- Goela P, Danchenko S, Icely J et al (2014) Using CHEMTAX to evaluate seasonal and interannual dynamics of the phytoplankton community off the South-west coast of Portugal. *Estuar Coast Shelf Sci* 151:112–123. <https://doi.org/10.1016/j.ecss.2014.10.001>

- Gokaraju B, Durbha SS, King RL et al (2011) A machine learning based spatio-temporal data mining approach for detection of harmful algal blooms in the Gulf of Mexico. *IEEE J Sel Topics Appl Earth Observ Remote Sens* 4(3):710–720. <https://doi.org/10.1109/JSTARS.2010.2103927>
- Hardison DR, Holland WC, Currier RD et al (2019) HABscope: a tool for use by citizen scientists to facilitate early warning of respiratory irritation caused by toxic blooms of *Karenia brevis*. *PLoS One* 14(6):e0218489. <https://doi.org/10.1371/journal.pone.0218489>
- Harley JR, Lanphier K, Kennedy EG et al (2020) The Southeast Alaska Tribal Ocean Research (SEATOR) partnership: addressing data gaps in harmful algal bloom monitoring and shellfish safety in Southeast Alaska. *Toxins* 12(6):407. <https://doi.org/10.3390/toxins12060407>
- He H, Ma Y (2013) Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons. <https://doi.org/10.1002/9781118646106>
- Hirzel M, Kate K, Shinnar A et al (2019) Type-driven automated learning with LALE. *CoRR* <https://arxiv.org/abs/1906.03957>. arXiv:1906.03957v1
- Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. *Neural Netw* 2(5):359–366. [https://doi.org/10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8)
- Icely J, Fragoso B (2023) Potential effects of climate change on offshore aquaculture of the mediterranean mussel (*Mytilus galloprovincialis*, Lamarck, 1819) in Portugal. *J Shellfish Res* 42(2):223–235
- James SC, Zhang Y, O'Donncha F (2018) A machine learning framework to forecast wave conditions. *Coastal Eng* 137:1–10. <https://doi.org/10.1016/j.coastaleng.2018.03.004>
- Karlson B, Andersen P, Arneborg L et al (2021) Harmful algal blooms and their effects in coastal seas of Northern Europe. *Harmful Algae* 102:101989. <https://doi.org/10.1016/j.hal.2021.101989>
- Kavanaugh KE, Derner K, Fisher KM et al (2013) Assessment of the eastern Gulf of Mexico Harmful Algal Bloom Operational Forecast System (GOMX HAB-OFS): a comparative analysis of forecast skill and utilization from October 1, 2004 to April 30, 2008. Tech. rep., NOAA. <https://repository.library.noaa.gov/view/noaa/14750>
- Krug LA, Platt T, Sathyendranath S et al (2017) Unravelling region-specific environmental drivers of phytoplankton across a complex marine domain (off SW Iberia). *Remote Sens Environ* 203:162–184. <https://doi.org/10.1016/j.rse.2017.05.029>
- Kuhn M, Johnson K et al (2013) Applied predictive modeling, vol 26. Springer. <https://doi.org/10.1007/978-1-4614-6849-3>
- Kyriakides G, Margaritis KG (2019) Hands-on ensemble learning with Python: build highly optimized ensemble machine learning models using scikit-learn and Keras. Packt Publishing Ltd
- Langbridge A, O'Donncha F, Ba A et al (2023) Causal temporal graph convolutional neural networks (ctgcn). arXiv:2303.09634. <https://doi.org/10.48550/arXiv.2303.09634>
- Lary DJ, Müller MD, Mussa HY (2004) Using neural networks to describe tracer correlations. *Atmos Chem Phys* 4(1):143–146. <https://doi.org/10.5194/acp-4-143-2004>
- Lee S, Lee D (2018) Improved prediction of harmful algal blooms in four major South Korea's rivers using deep learning models. *Int J Environ Res Public Health* 15(7):1322. <https://doi.org/10.3390/ijerph15071322>
- Legleiter CJ, King TV, Carpenter KD et al (2022) Spectral mixture analysis for surveillance of harmful algal blooms (SMASH): a field-, laboratory-, and satellite-based approach to identifying cyanobacteria genera from remotely sensed data. *Remote Sens Environ* 279:113089. <https://doi.org/10.1016/j.rse.2022.113089>
- Leith C (1974) Theoretical skill of Monte Carlo forecasts. *Mon Weather Rev* 102(6):409–418. [https://doi.org/10.1175/1520-0493\(1974\)102<0409:TSOMCF>2.0.CO;2](https://doi.org/10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2)
- Loureiro S, Newton A, Icely JD (2005) Microplankton composition, production and upwelling dynamics in Sagres (SW Portugal) during the summer of 2001. *Sci Mar* 69(3):323–341. <https://doi.org/10.3989/scimar.2005.69n3323>
- McGillcuddy D Jr (2010) Models of harmful algal blooms: conceptual, empirical, and numerical approaches. *J Mar Sys* 83(3–4):105. <https://doi.org/10.1016/j.jmarsys.2010.06.008>
- O'Donncha F, Grant J (2019) Precision aquaculture. *IEEE Internet Things Mag* 2(4):26–30. <https://doi.org/10.1109/IOTM.0001.1900033>
- O'Donncha F, Hartnett M, Nash S et al (2015) Characterizing observed circulation patterns within a bay using HF radar and numerical model simulations. *J Mar Sys* 142:96–110. <https://doi.org/10.1016/j.jmarsys.2014.10.004>
- O'Donncha F, Hu Y, Palmes P et al (2022) A spatio-temporal LSTM model to forecast across multiple temporal and spatial scales. *Ecol Inf* 101687. <https://doi.org/10.1016/j.ecoinf.2022.101687>
- Oruche R, O'Donncha F (2023) Attention-based domain adaption forecasting of streamflow in data sparse regions. arXiv:2302.05386. <https://doi.org/10.48550/arXiv.2302.05386>
- O'Donncha F, Zhang Y, Chen B et al (2019) Ensemble model aggregation using a computationally lightweight machine-learning model to forecast ocean waves. *J Mar Syst* 199:103206. <https://doi.org/10.1016/j.jmarsys.2019.103206>

- O'Donncha F, Iakymchuk R, Akhriev A et al (2020) AllScale toolchain pilot applications: PDE based solvers using a parallel development environment. *Comput Phys Commun* 251:107089. <https://doi.org/10.1016/j.cpc.2019.107089>
- Park Y, Cho KH, Park J et al (2015) Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. *Sci Total Environ* 502:31–41. <https://doi.org/10.1016/j.scitotenv.2014.09.005>
- Pinto L, Mateus M, Silva A (2016) Modeling the transport pathways of harmful algal blooms in the Iberian coast. *Harmful Algae* 53:8–16. <https://doi.org/10.1016/j.hal.2015.12.001>
- Portuguese Institute for Sea and Atmosphere (2022) Coastal ocean forecasting. <https://www.ipma.pt/en/index.html>
- Portuguese Institute for Sea and Atmosphere (2023) Legal limits on bivalve molluscs. https://www.ipma.pt/pt/bivalves/docs/files/Limites_legais_em_moluscos_bivalves.pdf
- Provost F (2000) Machine learning from imbalanced data sets 101. In: *Proceedings of the AAAI'2000 Workshop on Imbalanced Datasets*, AAAI Press, pp 1–3. <http://hdl.handle.net/2451/27763>
- Roiha P, Westerlund A, Nummelin A et al (2010) Ensemble forecasting of harmful algal blooms in the Baltic Sea. *J Mar Syst* 83(3–4):210–220. <https://doi.org/10.1016/j.jmarsys.2010.02.015>
- Samek W, Montavon G, Vedaldi A et al (2019) Explainable AI: interpreting, explaining and visualizing deep learning, vol 11700. *Springer Cham*. <https://doi.org/10.1007/978-3-030-28954-6>
- Santos M, Moita MT, Oliveira PB et al (2021) Phytoplankton communities in two wide-open bays in the Iberian upwelling system. *J Sea Res* 167:101982. <https://doi.org/10.1016/j.seares.2020.101982>
- Shapley LS (1997) A value for n-person games. *Classics in game theory* 69. <https://doi.org/10.7249/P0295>
- Silva A, Pinto L, Rodrigues S et al (2016) A HAB warning system for shellfish harvesting in Portugal. *Harmful Algae* 53:33–39. <https://doi.org/10.1016/j.hal.2015.11.017>
- Tonani M, Sykes P, King RR et al (2019) The impact of a new high-resolution ocean model on the Met Office North-West European Shelf forecasting system. *Ocean Sci* 15:1133–1158. <https://doi.org/10.5194/os-15-1133-2019>
- Trainer VL (2020) GlobalHAB: evaluating, reducing and mitigating the cost of harmful Algal blooms: a compendium of case studies. *North Pacific Marine Science Organization*. <https://doi.org/10.25607/OBP-1709>
- Trainer VL, Hardy FJ (2015) Integrative monitoring of marine and freshwater harmful algae in Washington State for public health protection. *Toxins* 7(4):1206–1234. <https://doi.org/10.3390/toxins7041206>
- Turner M, Walker J, Oke P (2008) Ensemble member generation for sequential data assimilation. *Remote Sens Environ* 112(4):1421–1433. <https://doi.org/10.1016/j.rse.2007.02.042>
- Van Dolah FM (2000) Marine algal toxins: origins, health effects, and their increased occurrence. *Environ Health Perspect* 108(suppl 1):133–141. <https://doi.org/10.1289/ehp.00108s1133>
- Villali J (2021) The increasing importance of advanced weather technologies: building resiliency in utility operations and field services. *Tech. rep, IDC Technology Spotlight*
- Wolff S, O'Donncha F, Chen B (2020) Statistical and machine learning ensemble modelling to forecast sea surface temperature. *J Mar Syst* 208:103347. <https://doi.org/10.1016/j.jmarsys.2020.103347>
- Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. *IEEE Trans Evol Comput* 1(1):67–82. <https://doi.org/10.1109/4235.585893>
- Wynne TT, Stumpf RP, Tomlinson MC et al (2011) Estimating cyanobacterial bloom transport by coupling remotely sensed imagery and a hydrodynamic model. *Ecol Appl* 21(7):2709–2721. <https://doi.org/10.1890/10-1454.1>
- Wynne TT, Meredith A, Stumpf RP et al (2020) Harmful algal bloom forecasting branch ocean color satellite imagery processing guidelines, 2020 update. *Tech. Rep. NOAA technical memorandum NOS NCCOS ; 296, NOAA Technical Memorandum*. <https://doi.org/10.25923/606t-m243>
- Yang Y, Dong J, Sun X et al (2017) A CFCC-LSTM model for sea surface temperature prediction. *IEEE Geosci Remote Sens Lett* 15(2):207–211. <https://doi.org/10.1109/LGRS.2017.2780843>
- Zhang F, Wang Y, Cao M et al (2016) Deep-learning-based approach for prediction of algal blooms. *Sustainability* 8(10):1060. <https://doi.org/10.3390/su8101060>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.