



Highly explainable predictive models with DoME for the management of DSP-related harmful algal blooms in the shellfish industry

Andres Molares-Ulloa^{1,2} , Daniel Rivero^{1,2} , Jose Molares³ , Enrique Fernandez-Blanco^{1,2} 

¹ Universidade da Coruña, Department of Computer Science and Information Technology, Faculty of Computer Science, 15071, A Coruña, Spain

² Centro de investigación CITIC, Department of Computer Science and Information Technology, University of A Coruña, 15071, A Coruña, Spain

³ Instituto Tecnológico para o Control do Medio Mariño de Galicia (INTECMAR), Consellería do Mar. Peirao de Vilaxoán, s/n 36611, Vilagarcía de Arousa, Galicia, Spain.

Corresponding author: andres.molares@udc.es

Abstract: The occurrence of HAB has a direct impact on shellfish farming, leading to economic losses due to the contamination of shellfish with toxins harmful to human health. Predicting these blooms accurately is therefore crucial for minimizing their negative effects on the industry. The DoME machine learning model is particularly notable for its high interpretability, as the trained model is expressed as a mathematical equation, allowing for transparent analysis and a better understanding of the factors driving the predictions. This characteristic distinguishes DoME from other black-box models, making it a valuable tool for stakeholders seeking not only accurate predictions but also insights into the dynamics behind HAB events. In this study, we evaluated the novel DoME (Development of Mathematical Expressions) algorithm for the prediction of Harmful Algal Blooms (HAB) associated with Diarrhoeic Shellfish Poisoning (DSP), a significant concern for the shellfish industry. Our testing involved analysing the model's performance in various environmental conditions, demonstrating its robustness and adaptability. DoME achieved a F1-score of 97.80%, which corresponds to an improvement of around 8% over previous studies. This superior performance, combined with its explainability, underscores the model's potential as a practical and reliable solution for early warning systems in the shellfish industry, helping to protect both public health and economic stability.

Keywords: Machine Learning, Harmful Algal Blooms, Biotoxins, Aquaculture, Symbolic Regression.

1 Introduction

The production of shellfish is vital worldwide, not only for its economic value but also for its contribution to food security and the sustainability of coastal communities. Spain, in particular, plays a prominent role in this industry, being one of the largest producers and exporters of shellfish, especially from the region of Galicia. Production in this region accounts for around 40% of European production [1]. However, the shellfish industry faces significant risks due to Harmful Algal Blooms (HABs), which can produce harmful toxins [2]. HABs refer to the excessive proliferation of certain algal species that produce toxins harmful to marine life, aquatic ecosystems, and human health. In the Galician coast, the most common toxins

are marine biotoxins, including Amnesic Shellfish Poisoning (ASP), Paralytic Shellfish Poisoning (PSP), and Diarrhoeic Shellfish Poisoning (DSP), with the latter being the most common and thus having the greatest impact [3]. These toxins can accumulate in shellfish flesh, making their consumption dangerous and leading to the temporary closure of cultivation areas, resulting in significant economic losses for the industry [4, 5]. Therefore, effective monitoring and management of marine waters are essential to mitigate these risks.

There is significant scientific interest in understanding the causes and effects of the spatial and temporal distribution of HAB species due to their potential impacts on ecosystems, public health, tourism, and social structures, all of which result in substantial economic losses. Continuous monitoring is crucial to take preventive action when HABs appear. Although HABs are natural and cannot be prevented, active surveillance is necessary to monitor their occurrence and take action accordingly. Within the European Union, the management of mollusc production areas involves analysing toxicity in mussel meat [6, 7, 8]. When such analyses are not feasible, authorities rely on measurable factors that favour the proliferation of toxic phytoplankton to make decisions. Currently, production area closures are based on expert knowledge without predictive models, causing disruptions and economic losses in the industry. The existence of models to support decision-making could help reduce the impact on the industry by enabling the creation of contingency plans [9]. Recent studies indicate that machine learning techniques are the best alternative for creating these models [10, 11, 12].

There are two main approaches to creating these predictive models: one focuses on predicting the concentration of specific harmful cells [13, 14, 15, 16, 17], and the other on predicting biomarkers closely related to HABs, such as chlorophyll-a (chl-a) concentration or toxin levels [18, 19, 20, 21]. While these are the main approaches due to the way sampling is conducted, they present certain drawbacks in predicting HABs. These events can be caused by multiple phytoplankton species simultaneously, so predicting specific species may be insufficient. Chl-a is a biomarker indicating a high presence of plant mass in the water, but not all microalgae produce toxins, which can lead to erroneous models in regions where water eutrophication is not an issue. Predicting HABs based on the presence of toxins is less common due to irregular data from established sampling programs, but it provides a much more reliable objective variable for creating these systems. Based on this last idea, there is a less common option where the status of the production area is used as the target variable [22, 23], based on the toxicity level of shellfish meat relative to legal limits.

For predicting HABs, machine learning techniques such as Support Vector Machines (SVM), Multi-Layer Perceptrons (MLPs), Random Forests (RF), and Convolutional Neural Networks (CNN) are often among the most successful. SVMs are efficient in high-dimensional spaces and robust against overfitting, but they require extensive training time and are not intuitively interpretable. This model has been used by González Vilas et al. [15], showing good results in predicting ASP blooms caused by *Pseudo-nitzschia spp.* Additionally, the model developed by Xiu Li et al. [19] for predicting chl-a in Tolo Harbour, Hong Kong, was developed using this technique, although it offers the best results, its execution time is very high. RFs are robust and handle many independent variables well, but the combination of multiple trees can make detailed interpretation difficult. Studies like those by Hiroshi Yajima and Jonathan Derot [20], Jonathan Derot, Hiroshi Yajima, and Stéphan Jacquet [17], or John R. Harley et al. [24] position this model as the best alternative under certain conditions, such as predicting HABs in freshwater or certain species like *Planktothrix rubescens*. MLPs, while highly flexible and capable of modelling complex relationships, require large amounts of data, are prone to overfitting, and their interpretation is complex due to their multi-layered structure. Due to their versatility, this is one of the most studied techniques, applied in predicting chl-a [25], cell concentration of the genus *Skeletonema* [26], *Dinophysis acuminata* [13], *Pseudo-nitzschia* diatoms and *Karlodinium* dinoflagellates [16] as well as other biomarkers [27, 28]. CNNs, effective in identifying spatial and temporal patterns, have advanced all fields of study, including HAB prediction [29]. However, these networks require an input data structure in the form of maps, which is not always feasible due to the inherent limitations of sampling programs. Additionally, the need for large volumes of data, significant computational power, and their complex structure complicates result interpretation, posing a significant handicap in decision-making systems for HAB control. Other less common models have outperformed these techniques under certain conditions, such as Bootstrap Aggregated Network (BAGNET), a hybrid technique based on an ensemble of MLPs that achieves results of 90% in the Galician coast [30].

A common challenge in using these machine learning techniques is the explainability of the models. Complex models like ANN and CNN are often considered “black boxes” because the internal relationships leading to predictions are not easily interpretable. This can be problematic in applications requiring transparency and understanding of the decision-making process, such as risk management and environmental policy formulation. SVMs and RFs, although more interpretable than ANN and CNN, still present challenges in explaining the individual contributions of variables to the final predictions.

Bearing in mind the need for high explainability and the absence of a dominant model in this field of study, we have developed a highly explainable predictive model for harmful algal blooms and compared it with the most common models in the literature. Emphasizing the main contributions of our approach:

- The model targets the toxicity level of shellfish meat produced by DSP toxin.
- The technique to be used is DoME (Development of Mathematical Expressions) [31], an approach that allows the extraction of mathematical expressions from a dataset.
- This represents a significant advance both in the execution speed of the developed model and in the field of explainability, as DoME facilitates the interpretation of the underlying relationships in the data through clear and understandable mathematical expressions.

The proposed workflow for the DoME-based modelling approach is illustrated in Figure 1.

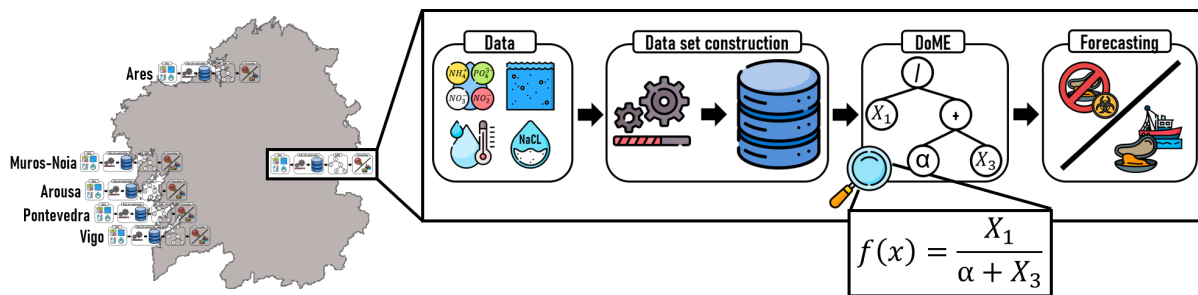


Figure 1: Workflow diagram illustrating the methodological steps of the proposed approach. The diagram outlines the sequential stages involved in data preprocessing, model construction using the DoME algorithm, expression optimization, and final evaluation.

2 Materials and Methods

2.1 Dataset and its construction

This study was conducted along the coast of Galicia (42.8°N 7.9°W). This region has a significant mollusc industry, being one of the most important in Spain. The data required for the study and model training were collected through various means. Most of these data were gathered by the *Instituto Tecnológico para o Control do Medio Mariño de Galicia* (INTECMAR) [32] across 42 oceanographic stations that are part of its HAB monitoring network (these stations can be seen in Figure 2). Additional data used in this study were obtained from the *marnaraia* project [33] of the *Instituto Español de Oceanografía* (IEO). The data used comes from samples collected between the years 2004 and 2019.

The samples were labelled based on the toxicity levels measured on mussel flesh sampled on the production areas seen in Figure 2. Table 1 shows the volume and distribution of the samples according to the estuary. Due to the heterogeneity of the data, it was decided to create independent models for each estuary under study. Analysing each estuary separately allows us to create data sets with varying levels of detail. For example, the Ares-Betanzos estuary is the simplest to analyse due to its smaller number of

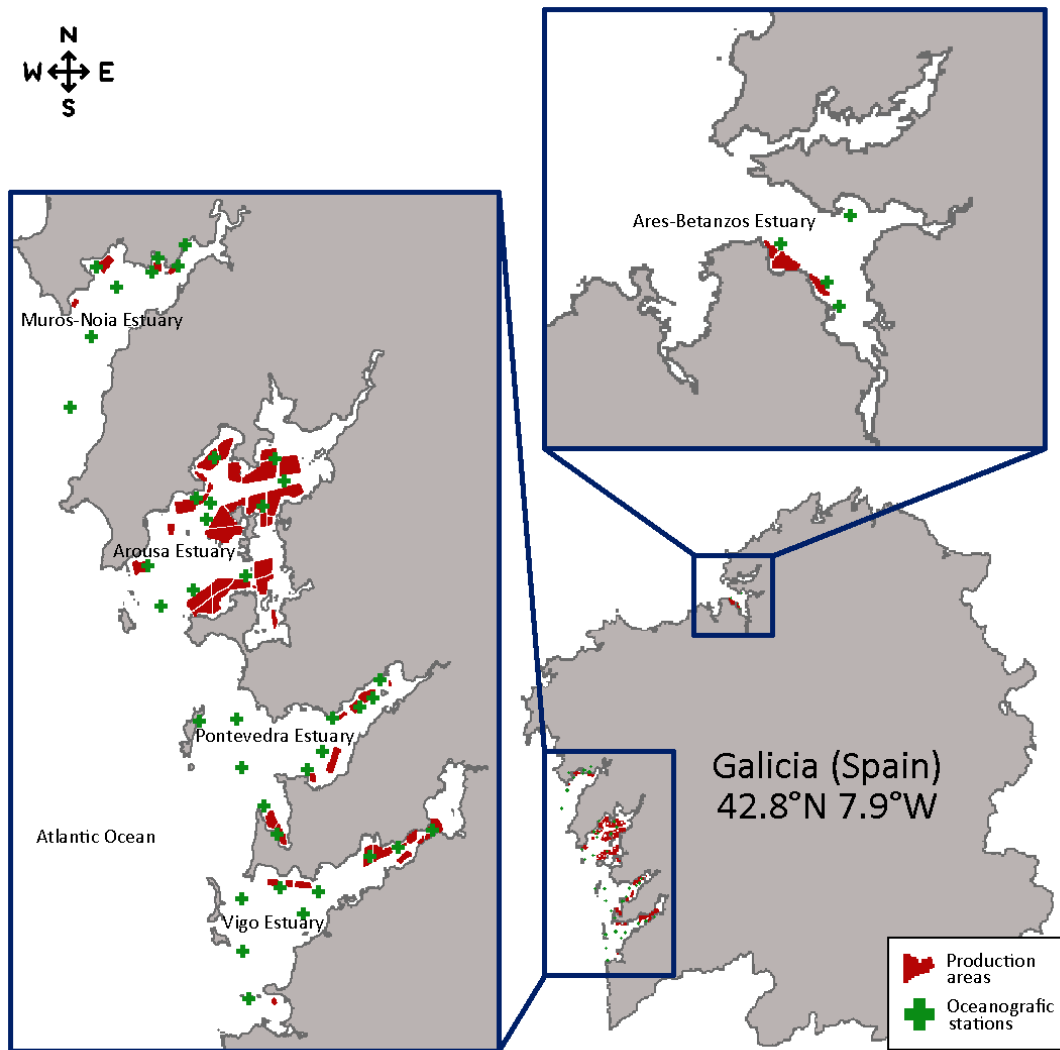


Figure 2: Map of Galicia and the location of the production areas studied marked in red and the oceanographic stations marked with green crosses.

stations and production areas (4 and 2, respectively). In contrast, the Arousa estuary is more complex, with 10 stations and 22 production areas.

Initially, the possibility of treating the data as a time series was considered, taking into account the evolution of variables over time. However, due to the nature of the data, characterized by a weekly sampling frequency, a weekly dataset approach was chosen instead. These datasets were labelled based on the presence of toxins on the Monday following the sampling date, as decisions regarding the opening and closing of production areas on Mondays are made without the availability of recent analytical data (the sampling program stops on weekends).

The variables used for training the models range from data on concentrations of photosynthetic pigments and the abundance of microalgal species to physico-chemical parameters such as nutrients, temperature, and salinity, as well as derived oceanographic variables. These features were processed and adapted to create the datasets in the following way:

- Chlorophyll: The maximum concentrations of chlorophyll a, b, and c were used.
- Dinophysis: The count of various Dinophysis species (*Dinophysis acuminata*, *Dinophysis acuta*,

	Ares-Betanzos	Muros-Noia	Arousa	Pontevedra	Vigo
No. oceanographic stations	4	8	10	11	9
No. production areas	2	4	24	8	12
Samples	1,564	3,128	18,768	6,256	9,384
Samples without null values	558	1,440	12,168	3,760	5,112
Closures without null values	193 (35%)	508 (35%)	1903 (16%)	1858 (49%)	1168 (23%)
Openings without null values	365 (65%)	932 (65%)	10265 (84%)	1902 (51%)	3944 (77%)

Table 1: Distribution of oceanographic stations, production areas and openings and closures in each estuary

Dinophysis caudata and *Dinophysis spp.*), known producers of DSP toxin, was included.

- **Nutrients:** The concentration values of dissolved phosphate, nitrate, nitrite, and ammonium were considered.
- **Hydrographic Parameters:** Average values of temperature, salinity, and oxygen were employed. Additionally, thermocline and halocline stratification were assessed by calculating the absolute difference between the mean temperature and salinity values in the top 6 meters and the next 6 meters.
- **Upwelling Index:** The weekly mean value of the upwelling index was used.
- **Production Area Status:** Production areas were classified as open or closed based on their toxicity levels relative to the legal limit. This classification served as both an output and input parameter. The output parameter was the Monday value of the week following the study week, while the input parameter was the Friday value, the closest sampling day to the prediction date.
- **Production Area Encoding:** The production area of each sample was encoded using one-hot encoding [34].
- **Sampling Date:** The sampling date was calculated as a sine function.

Five datasets were created, one for each estuary. The datasets contained a substantial number of data inconsistencies, potentially due to technical issues, sampling difficulties, or the late establishment of certain stations. To ensure model compatibility, samples with missing values were removed, resulting in the distribution of openings and closings shown in Table 1.

2.2 DoME

DoME is an algorithm designed to solve regression problems by creating an equation that captures the relationship between independent variables and a dependent variable [31]. This process, where the goal is to derive an equation as the model, is referred to as Symbolic Regression. Additionally, DoME can handle classification tasks. For binary classification, the algorithm incorporates the Heaviside step function into the output, enabling it to distinguish between two classes.

Within DoME, the equations are structured as trees, composed of terminal and non-terminal nodes. The terminal nodes consist of the independent variables and constants, while non-terminal nodes represent operators that contribute to the equation. The operators in this algorithm are basic arithmetic functions (+, -, *, /), and the complexity of the resulting expressions can be constrained by limiting the number of nodes in the tree.

The algorithm starts with an initial tree, where a constant equal to the average of the dependent variable is used. An iterative process then follows, in which the tree undergoes gradual modifications during each cycle, continuing until no further changes lead to better predictions. Figure 3 presents a diagram of the DoME model throughout its training phase.

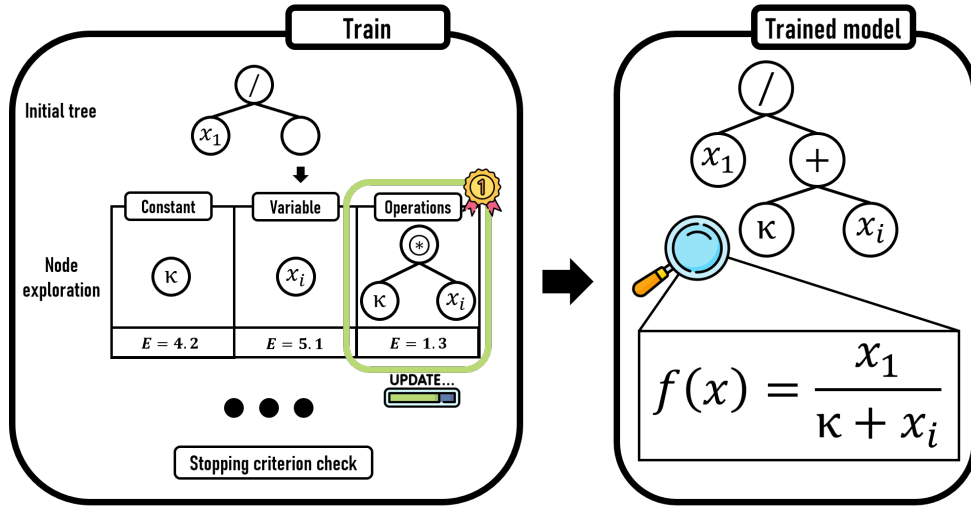


Figure 3: Flow diagram of the DoME algorithm. The process begins with an initial constant-based expression tree, which is iteratively expanded by evaluating possible node substitutions—using constants, variables, or arithmetic expressions—based on error reduction. Optimal replacements are selected if they significantly improve performance, and the procedure repeats until a stopping criterion is satisfied.

2.3 Performance measures

To assess and compare the trained models, four key performance metrics were considered: accuracy, recall, precision and F1-score. In the confusion matrix used for these calculations, closures of production zones were defined as positive, while openings were defined as negative. A brief description of the metrics used is shown in Table 2.

Metric	Equation	Description
Accuracy	$\frac{TP+TN}{TP+FP+FN+TN}$	The overall proportion of correct predictions out of the total number of predictions.
Recall	$\frac{TP}{TP+FN}$	The proportion of positive instances that were correctly identified out of the total number of actual positive instances.
Precision	$\frac{TP}{TP+FP}$	The proportion of positive predictions that were actually correct out of the total number of positive predictions.
F1 – score	$\frac{2 \cdot (\text{Recall} \cdot \text{Precision})}{(\text{Recall} + \text{Precision})}$	The harmonic mean of precision and recall, providing a balanced measure of both.

Table 2: Equations and brief description of the metrics used in them

Given the uneven distribution of data, traditional metrics like accuracy may not properly reflect model performance. Consequently, a combination of accuracy, F1-score, and, most importantly, recall will be considered. Prioritizing recall is crucial to minimize false negatives, as misclassifying a potential closure as an open production area could pose a significant risk to public health.

2.4 Experimentation setup

To enhance model reliability, we applied K-fold cross-validation [35], specifically 10-fold cross-validation. This method is a technique used to evaluate the performance of a machine learning model in a more robust manner. The process involves dividing the dataset into 10 folds of approximately equal size. In

each iteration, 9 of these folds are used to train the model, while the remaining fold is used to evaluate it. This procedure is repeated 10 times, changing which fold is used for evaluation in each iteration. In the end, the average of the evaluation metrics obtained in each iteration is calculated to provide a more reliable estimate of the model’s performance. Given the imbalanced nature of the data, stratified K-fold was used to ensure balanced folds.

Model hyperparameters were optimized using grid search, an exhaustive method that evaluates various combinations of hyperparameter values. These values were determined through empirical experimentation. The specific grid search configuration for the model is presented in Table 3.

DoME	
Minimum Reductions MSE	$1e^{-1}$, $1e^{-2}$, $1e^{-3}$, $1e^{-4}$, $1e^{-5}$, $1e^{-6}$ and $1e^{-7}$
MaxNumNodes	5:5:150
Strategy	“Exhaustive with constant optimization”, “Selective” and “Selective with constant optimization”

Table 3: Parameter values used in the grid search.

3 Results and Discussion

Once the different models (one for each estuary) were trained and validated using the 10-fold technique, a series of metrics were obtained. These metrics are shown in Figure 4, where we can observe the mean and standard deviation of the studied metrics. Additionally, we can see the model’s behaviour depending on the estuary where it was trained. Notably, the overall results are strong, especially in estuaries like Ares-Betanzos and Vigo. When training these models, it is noteworthy that those trained on the Ares-Betanzos and Pontevedra estuaries provided the best results with the “Selective with constant optimization” strategy, while in the other estuaries, the best configuration used the “Selective” strategy.

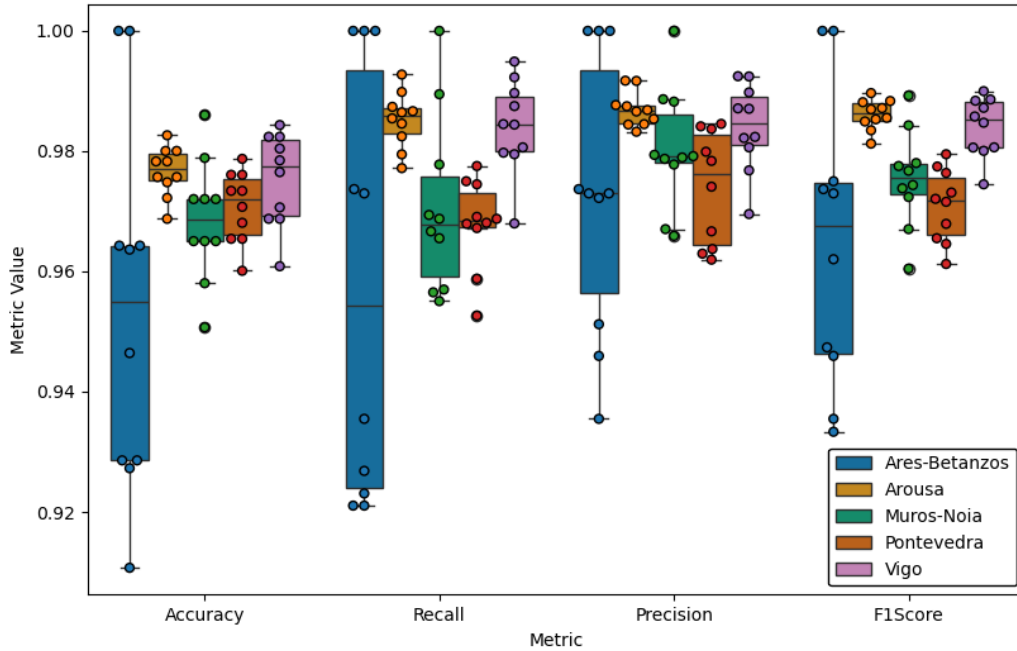


Figure 4: Comparison of accuracy, recall, precision and F1-score for the prediction of HAB episodes by DSP by DoME. The model was tested in different datasets belonging to 5 estuaries.

We made an analysis based on the model comparative shown in Table 4, which includes widely used models such as Random Forest (RF), Artificial Neural Networks (ANN), k-Nearest Neighbors (kNN), Support Vector Machines (SVM), XGBoost, and BAGNET. Lastly, an in-depth analysis of the DoME model is provided, which demonstrates superior performance across all evaluated metrics, both in terms of mean and standard deviation.

With an accuracy of 96.89%, DoME achieved the best results in addressing this specific problem. Although the improvement in recall compared to other techniques was about 4%, reaching 97.30%, the F1-score was increased by 8%, reaching 97.80%, highlighting the significant imbalance between recall and precision in other models. While recall is the most important metric due to the high impact of false negatives, false positives can also result in significant losses for the industry.

MODEL	ACCURACY		RECALL		F1-SCORE		SOURCE
RF	94.15	±2.08	88.75	±7.19	89.23	±5.66	[24]
ANN	91.36	±6.01	89.93	±6.86	85.51	±11.59	[28]
kNN	92.78	±2.01	86.80	±5.76	87.01	±4.98	[23]
SVM	92.97	±4.59	84.00	±17.76	85.98	±11.72	[19]
XGBoost	92.34	±5.13	82.81	±15.81	85.30	±11.25	[36]
BAGNET	93.75	±3.20	93.41	±3.61	89.27	±6.94	[30]
DoME	96.89	±0.83	97.30	±1.05	97.80	±0.62	Proposed here

Table 4: Metrics of each model averaged in the 5 estuaries.

Thanks to the nature of the DoME algorithm, the trained model consists of a mathematical equation that can be easily applied to input features in any calculation software. This makes DoME an extremely fast and easy-to-use tool for predicting harmful algal blooms (HAB) in the aquaculture industry. As an example of this, Eq. 1 shows the trained model for predicting the status of production areas affected by

DSP HABs in the Ares-Betanzos estuary.

$$State = \mathcal{U} \left(1.73374 \cdot \left(\left(-0.00089 \cdot \left(T_0 - \left(2.61222 \cdot X_{12} + \frac{0.53734 + X_{21} + T_1}{X_{14}} - 1.97190 \right) \cdot X_3 \right) \cdot X_{52} \right) + X_{69} \right) \right) \quad (1)$$

Where T_0 is calculated as Eq. 2 and T_1 is calculated as Eq. 3.

$$T_0 = \frac{(-0.00532 \cdot X_{69} \cdot X_{24} + (0.00514 \cdot X_{49})) \cdot X_{36} + (36.68607 - X_{22}) \cdot X_{49}}{X_{49}} \quad (2)$$

$$T_1 = \frac{(X_{11} + X_4 - 0.92492) \cdot (1.07237 + X_{24}) \cdot X_{14}}{\left(\frac{-0.02570}{X_{26} \cdot X_{14}} \cdot (-123.92966 \cdot X_{39} + X_{43} + 0.02120) \right) + 0.22198 - X_{18}} \quad (3)$$

The variables defined as X_i , where i represent an internal index, represent the features showed in Table 5. Out of the many initial variables used to train the algorithm. DoME found the best performance by using only those present in Table 5. The missing values of i correspond to the variables that were not used for the final algorithm.

Variable	Feature
X_3	Production area
X_4	Ammonium dissolved in water measured in sampling area L1
X_{11}	Nitrite dissolved in water measured in sampling area L2
X_{12}	Ammonium dissolved in water measured in sampling area L3
X_{14}	Nitrate dissolved in water measured in sampling area L3
X_{18}	Nitrate dissolved in water measured in sampling area L4
X_{21}	Chlorophyll-b measured in sampling area L1
X_{22}	Chlorophyll-c measured in sampling area L1
X_{24}	Chlorophyll-b measured in sampling area L2
X_{36}	Concentration of <i>D. acuminata</i> measured in sampling area L2
X_{39}	Concentration of <i>D. spp</i> measured in sampling area L2
X_{43}	Concentration of <i>D. spp</i> measured in sampling area L3
X_{49}	Halocline stratification index measured in sampling area L1
X_{52}	Water temperature measured in sampling area L1
X_{69}	Opening status of the production area on previous Friday

Table 5: Relationship between variables and features in the equations shown in the study.

When comparing these results with previous studies, one of the most notable aspects is the reduction of the standard deviation across the different geographical regions where the models were applied. This demonstrates that DoME is a robust model with a remarkable capacity for adaptation, performing effectively even in areas with a significant imbalance between negative and positive instances in the data, such as the Ría de Arousa.

This advantage of highly explainable models has been underexplored in this field. Although algorithms like kNN or Random Forest are often preferred for their interpretability, it has been shown that their performance is inferior compared to the proposed one.

4 Conclusions

In this article, we trained a machine learning model using the DoME algorithm for the prediction of HABs. The model's performance was analysed in five different locations, each with distinct environmental characteristics and ecosystem conditions. This allowed us to observe how the model adapts to various ecological contexts, showing variations in accuracy depending on local factors, which highlights its robustness and generalization capability. We examined the performance of various machine learning algorithms and compared them with DoME, evaluating their accuracy, recall, and F1-score, as we can see in Table 4. Each of these metrics reflects a specific aspect of the models' performance, providing a

comprehensive view of their predictive capabilities. Achieving performance above 97% across all metrics makes it an excellent option for predicting HABs in the aquaculture industry.

Inadequate preparation for these natural phenomena can lead to substantial economic losses for aquaculture and shellfish industries. In this context, the DoME algorithm, proposed in this study, has the potential to be a valuable tool for producers, as it could allow them to anticipate possible closures in production areas several days in advance, optimizing both resource management and the efforts of monitoring services during HAB episodes.

Unlike other models, DoME based models maintained a much more appropriate balance between recall and precision. This balance is crucial in contexts where both false negatives and false positives can have significant impacts. DoME showed a notable improvement in recall, which is essential for minimizing the risks associated with failing to detect HABs.

One of the most remarkable features of the DoME model was its ability to adapt to different geographic regions, with a significant reduction in the standard deviation of its performance. This suggests that the model is highly robust, performing effectively even in areas with imbalanced data characteristics. Despite the model's adaptability across regions, it would be beneficial to study how to fine-tune its performance in geographic areas with extreme environmental conditions or highly sparse data. This could strengthen its accuracy in even more challenging environments.

It is crucial to highlight the importance of developing highly interpretable models, which not only offer outstanding performance but also enable decision-makers to better understand the factors influencing these phenomena, facilitating a faster and more efficient response to potential HAB events. The integration of DoME into real-time monitoring platforms could be a natural extension. While this depends on the existence of such platforms, it would allow the creation of automated early warning systems that send alerts to aquaculture producers when conditions conducive to HABs are detected, enabling a more effective preventive response.

Although DoME was originally conceived as a regression algorithm, this article demonstrates its potential as a classification technique. The results obtained indicate that DoME can not only adapt to classification problems but also improve performance compared to other conventional classification algorithms. This suggests that DoME offers significant versatility and opens up new opportunities for its application in various machine learning tasks.

This algorithm integrates a feature selection process that allows it to handle large volumes of features. This characteristic could be used independently in combination with other classification algorithms to create a workflow where DoME is used to select the most important features for training a classifier. This classifier could be based on DoME or any other method.

Acknowledgements

The authors want to acknowledge the support from INTECMAR, who has provided most of the data for this work; and CESGA, who allowed the conduction of the tests on their installations. CITIC is funded by the Xunta de Galicia through the collaboration agreement between the Regional Ministry of Culture, Education, Vocational Training and Universities and the Galician universities to strengthen the research centres of the Galician University System (CIGUS). Grant PID2021-126289OA-I00 funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe.

References

- [1] Apromar. La acuicultura en españa 2022. https://apromar.es/wp-content/uploads/2022/10/La_acuicultura_Espana_2022_v3_APROMAR.pdf, 27 March 2022.
- [2] Joann M. Burkholder. Implications of harmful microalgae and heterotrophic dinoflagellates in management of sustainable marine fisheries. *Ecological Applications*, 8(sp1):S37–S62, 1998.
- [3] F. Vilas, Daniel Rey, Belén Rubio Armesto, Ana Bernabéu, Gonzalo Méndez, Ruth Durán, K. Mohamed, Gabriel Rosón, J. M. Cabanas, Fiz F. Pérez, Carmen G. Castro, Aida F. Ríos, F. G.

- Figueiras, Ana Miranda, Isabel Riveiro, Alba Ruth Vergara, C. Guisande, B. Reguera, Laura Escalera, Yolanda Pazos, Ángeles Moroño, Juan José González, Cristina Álvarez, Ricardo Beiras, Victoria Besada, J. Fumega, María Ángeles Franco, Mariano Gómez, Amelia González Quijano, Teresa Nunes, R. Prego, Antonio Soriano Sanz, Lucía Viñas, J. B. Peleteiro, Valentín Trujillo, Rafael Bañón, Jorge Ribó, M. Olmedo, Blanca Álvarez Blázquez, José Luis Rodríguez, Juan Pazó, Juan José Otero, Ángel Guerra, Santiago Lens, Francisco Rocha, María Xosé Vázquez Rodríguez, and Albino Prada Blanco. La ría de vigo : una aproximación integral al ecosistema marino de la ría de vigo. <http://hdl.handle.net/10261/170032>, 2008.
- [4] Hilary Corlett and Brian Jones. Epiphyte communities on thalassia testudinum from grand cayman, british west indies: Their composition, structure, and contribution to lagoonal sediments. *Sedimentary Geology*, 194(3):245–262, 2007.
- [5] Lucie Maranda, Susannah Corwin, Stacie Dover, and Steve L. Morton. Procoentrum lima (dinophyceae) in northeastern usa coastal waters: Ii: Toxin load in the epibiota and in shellfish. *Harmful Algae*, 6(5):632–641, 2007.
- [6] Reglamento (ce) n^o 853/2004 del parlamento europeo y del consejo de 29 de abril de 2004 por el que se establecen normas específicas de higiene de los alimentos de origen animal. <https://www.boe.es/doue/2004/139/L00055-00205.pdf>, 2004.
- [7] Reglamento (ue) n^o 786/2013 de la comisión, de 16 de agosto de 2013 , por el que se modifica el anexo iii del reglamento (ce) n^o 853/2004 del parlamento europeo y del consejo en lo relativo a los límites autorizados de yesotoxinas en moluscos bivalvos vivos. <http://data.europa.eu/eli/reg/2013/786/oj>, 2013.
- [8] Commission implementing regulation (eu) 2019/627 of 15 march 2019 laying down uniform practical arrangements for the performance of official controls on products of animal origin intended for human consumption in accordance with regulation (eu) 2017/625 of the european parliament and of the council and amending commission regulation (ec) no 2074/2005 as regards official controls. http://data.europa.eu/eli/reg_impl/2019/627/2021-01-01, 2019.
- [9] Di Jin and Porter Hoagland. The value of harmful algal bloom predictions to the nearshore commercial shellfish fishery in the gulf of maine. *Harmful Algae*, 7, 2008.
- [10] Rafaela C. Cruz, Pedro Reis Costa, Susana Vinga, Ludwig Krippahl, and Marta B. Lopes. A review of recent machine learning advances for forecasting harmful algal blooms and shellfish contamination. *Journal of Marine Science and Engineering*, 9, 2021.
- [11] Ayesha Aslam, Kashif Naseer Qureshi, and Adil Hussain. Advanced ai-based healthcare systems applications and services. In *Next Generation AI Language Models in Research*, pages 86–113. CRC Press, 2024.
- [12] Adil Hussain, Peng-Cheng Xu, Wang Shixin, and Kashif Naseer Qureshi. Artificial intelligence in natural science research: Innovations and limitations. In *Next Generation AI Language Models in Research*, pages 129–169. CRC Press, 2024.
- [13] L. Velo-Suárez and J. C. Gutiérrez-Estrada. Artificial neural network approaches to one-step weekly prediction of dinophysis acuminata blooms in huelva (western andalucía, spain). *Harmful Algae*, 6, 2007.
- [14] VH Aguilar Calderon. Predicción de las floraciones algales nocivas (fan) en poblaciones de dinophysis acuminata por redes neuronales artificiales. <https://repositorioslatinoamericanos.uchile.cl/handle/2250/3276312>, 2017.
- [15] Luis González Vilas, Evangelos Spyarakos, Jesus M. Torres Palenzuela, and Yolanda Pazos. Support vector machine-based method for predicting pseudo-nitzschia spp. blooms in coastal waters (galician rias, nw spain). *Progress in Oceanography*, 124, 2014.

- [16] Carles Guallar, Maximino Delgado, Jorge Diogène, and Margarita Fernández-Tejedor. Artificial neural network approach to population dynamics of harmful algal blooms in alfacas bay (nw mediterranean): Case studies of karlodium and pseudo-nitzschia. *Ecological Modelling*, 338, 2016.
- [17] Jonathan Derot, Hiroshi Yajima, and Stéphan Jacquet. Advances in forecasting harmful algal blooms using machine learning models: A case study with planktothrix rubescens in lake geneva. *Harmful Algae*, 99:101906, 2020.
- [18] Ashfaqur Rahman and Md Sumon Shahriar. Algae growth prediction through identification of influential environmental variables: A machine learning approach. *International Journal of Computational Intelligence and Applications*, 12, 2013.
- [19] Xiu Li, Jin Yu, Zhuo Jia, and Jingdong Song. Harmful algal blooms prediction with machine learning models in tolo harbour. In *2014 International Conference on Smart Computing*, pages 245–250, Nov 2014.
- [20] Hiroshi Yajima and Jonathan Derot. Application of the Random Forest model for chlorophyll-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases. *Journal of Hydroinformatics*, 20(1):206–220, 11 2017.
- [21] Xiaoyao Sun, Danyang Yan, Sensen Wu, Yijun Chen, Jin Qi, and Zhenhong Du. Enhanced forecasting of chlorophyll-a concentration in coastal waters through integration of fourier analysis and transformer networks. *Water Research*, page 122160, 2024.
- [22] Andrés Molares, Enrique Fernandez-Blanco, and Daniel Rivero. Application of artificial neural networks for the monitoring of episodes of high toxicity by dsp in mussel production areas in galicia. *Proceedings*, 54, 2020.
- [23] Andres Molares-Ulloa, Enrique Fernandez-Blanco, Alejandro Pazos, and Daniel Rivero. Machine learning in management of precautionary closures caused by lipophilic biotoxins. *Computers and Electronics in Agriculture*, 197:106956, 2022.
- [24] John R. Harley, Kari Lanphier, Esther Kennedy, Chris Whitehead, and Allison Bidlack. Random forest classification to determine environmental drivers and forecast paralytic shellfish toxins in southeast alaska with high temporal resolution. *Harmful Algae*, 99:101918, 2020.
- [25] Friedrich Recknagel, Jason Bobbin, Peter Whigham, and Hugh Wilson. Comparative application of artificial neural networks and genetic algorithms for multivariate time-series modelling of algal blooms in freshwater lakes. *Journal of Hydroinformatics*, 4(2):125–133, 03 2002.
- [26] Joseph H.W. Lee, Yan Huang, Mike Dickman, and A.W. Jayawardena. Neural network modelling of coastal algal blooms. *Ecological Modelling*, 159(2):179–201, 2003.
- [27] Isabella Grasso, Stephen D. Archer, Craig Burnell, Benjamin Tupper, Carlton Rauschenberg, Kohl Kanwit, and Nicholas R. Record. The hunt for red tides: Deep learning algorithm forecasts shellfish toxicity at site scales in coastal maine. *Ecosphere*, 10(12):e02960, 2019.
- [28] Jiu hao Guo, Yahong Dong, and Joseph H.W. Lee. A real time data driven algal bloom risk forecast system for mariculture management. *Marine Pollution Bulletin*, 161:111731, 2020.
- [29] Paul R. Hill, Anurag Kumar, Marouane Temimi, and David R. Bull. Habnet: Machine learning, remote sensing-based detection of harmful algal blooms. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:3229–3239, 2020.
- [30] Andres Molares-Ulloa, Daniel Rivero, Jesús Gil Ruiz, Enrique Fernandez-Blanco, and Luis de-la Fuente-Valentín. Hybrid machine learning techniques in the management of harmful algal blooms impact. *Computers and Electronics in Agriculture*, 211:107988, 2023.

-
- [31] Daniel Rivero, Enrique Fernandez-Blanco, and Alejandro Pazos. Dome: A deterministic technique for equation development and symbolic regression. *Expert Systems with Applications*, 198:116712, 2022.
- [32] INTECMAR. Web page of instituto tecnol3xico para o control do medio mari3o de galicia. <http://www.intecmar.gal/intecmar/default.aspx>, 22 January 2022.
- [33] IEO. Marnaraia proyect. <http://www.indicedeafloramiento.ieo.es/afloramiento.html>, 22 January 2024.
- [34] Pau Rodr3guez, Miguel A. Bautista, Jordi Gonz3lez, and Sergio Escalera. Beyond one-hot encoding: Lower dimensional target embedding. *Image and Vision Computing*, 75, 2018.
- [35] Tzu Tsung Wong. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48, 2015.
- [36] Moein Izadi, Mohamed Sultan, Racha El Kadiri, Amin Ghannadi, and Karem Abdelmohsen. A remote sensing and machine learning-based approach to forecast the onset of harmful algal bloom. *Remote Sensing*, 13(19), 2021.