Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 18(5): 6484–6505.
DOI: 10.3934/mbe.2021322
Received: 26 May 2021
Accepted: 19 July 2021
Published: 28 July 2021

Research article

Drivers of harmful algal blooms in coastal areas of Eastern Mediterranean: a machine learning methodological approach

Androniki Tamvakis, George Tsirtsis, Michael Karydis, Kleanthis Patsidis and Giorgos D. Kokkoris*

Department of Marine Sciences, Faculty of Environment, University of the Aegean, University Hill, GR81100, Mytilene, Greece

*** Correspondence:** Email: gkok@aegean.gr.

Abstract: Harmful algal species are present in the Mediterranean Sea and are often associated with toxic events affecting the nearby coastal zones. The presence of 18 marine microalgae, at genus level, associated with potentially harmful characteristics was predicted using a number of machine learning techniques based exclusively on a small set of abiotic variables, already identified as drivers of blooms. Random Forest (RF) algorithm achieved the best predictive performance by correctly identifying the presence of most genera with a mean of 89.2% of total samples. Although, RF has shown lower predictive performance for genera present in a low number of samples, its predictive power remains at least "fair' in these cases. The main tree-based advantage of RF was thereafter used to assess the importance of the input variables in predicting the presence of the algal genera. Temperature had the most powerful effect on genera's presences, although this effect varies among genera. Finally, the genera were clustered based on their response to the considered abiotic variables and common trends in an ecological context were identified.

Keywords: harmful algal; machine learning; Random Forest; abiotic parameters; Eastern Mediterranean

1. Introduction

Machine Learning (ML) algorithms are considered as powerful and reliable tools for prediction, applied so far in many scientific disciplines [1]. ML is a collection of algorithms of different types that improve prediction by gaining knowledge from data [2]. The main advantage of ML techniques is their ability to model multidimensional interactions between nonlinear, noisy, correlated or partly missing data. The field of marine ecology, characterized by complex interactions among a large number of variables, is a specific field counting several applications of ML algorithms [3].

Phytoplankton communities usually follow a rather typical annual pattern on biomass variations and species succession [4]. However, under a combination of forcing factors, it is possible that a regime shift in the system may be established. These factors can be the hydrodynamic conditions in the area [5], the buffering capacity of the system [6], salinity, temperature, nutrients, organic compounds [7], the presence of zooplankton and bacterial populations as well as species interactions [8]. However, work on those aspects [9], has shown that regime shifts are closely connected with the presence of nutrients; high sensitivity was also found in temperature fluctuations. These regime shifts in the case of phytoplankton are expressed as algal blooms. In many cases these blooms are due to exponential growth and dominance of toxic species of phytoplankton [10]. As the shift from the pristine conditions to an undesirable bloom does not follow a linear pattern, is of utmost importance to develop a probabilistic approach so that possible harmful effects on shellfish and fish culture as well as on human health can be indicated.

Algal blooms seem to be a problem of great concern among scientists due to the considerable economic, health and ecological effects on the neighboring coastal zones [11,12]. The various causes of blooms, often linked with eutrophication in coastal ecosystems, is a matter under consideration [13,14]. Many methods have therefore been applied aiming at associating different parameters expressing land use [15], human activities, nutrient loadings, climate change~~s~~ or other environmental processes with excessive algal growth [13,16,17]. These include the use of ML algorithms. In most cases, ML was applied to model and predict primary production (in terms of chlorophyll-α or biomass) using physical and chemical variables, as temperature, salinity, inorganic nitrogen and phosphorus concentrations [18–20]. In addition, there are studies that have modelled different dimensions of primary production such as diversity [21,22] or algal growth time [23]. A recent attempt has been made by Yu et al. (2021) [24] to predict phytoplankton blooms, using Machine Learning algorithms. Their argument is that "*when planktonic algae proliferate over certain limits, harmful algal blooms (HABs) will occur*". This is not necessarily the case [25]; increased concentrations of toxic algae can occur without a bloom and vice versa: not all blooms end to be toxic. Although algal blooms are entirely natural phenomena and have appeared throughout historical times [10], the frequency of occurrence has increased at an alarming level, over the last few decades. Land based sources of eutrophication, even when they are not the main drivers for triggering mechanisms for bloom formation, it is beyond any doubt that input of nutrients into the marine environment supports bloom formation. However, the principal question remains: will toxic microalgae grow? The answer can be given only if any predictions based on Machine Learning techniques, are targeted towards the exponential growth of toxic species, already recorded in the areas under study.

HABs are being considered as a specific form of coastal eutrophication, causing serious impact including death of organisms, toxins production, hypoxia and often human poisoning due to the consumption of infected fish or shellfish [14,26,27]. ML algorithms have been used to assess the formation of HABs in both fresh and marine waters [28]. Abundance as well as presence/absence of specific harmful algal species have been studied aiming to reveal possible associations with several biotic and abiotic variables. HAB forming species recently studied with selected ML techniques include *Karenia brevis* [29], *Planktothrix rubescens* [30] and *Dinophysis acuminata* [31], while

Bourel et al. [32] have used ensembles i.e., combinations of classifiers to predict the presence-absence of 8 harmful marine phytoplankton species. Considering the various ML algorithms, special attention has been given to Neural Networks (NNs) and Random Forests (RFs), two commonly used algorithms with great performance even when dealing with complex ecological processes as HAB formation [33–35]. Some applications of NNs include the assessment of algal population dynamics [36–38], the classification of different algal types [39,40] and detection of HABs from satellite images [41,42]. On the other hand, the increased effectiveness of RFs has been denoted on the assessment of the distribution of phytoplankton cell size [43] and on the efficient prediction of chlorophyll-*a* concentration [44,45], of the abundance of some toxic genotypes [46] and of the presence of some toxic algal species [47].

Variables used as drivers in the application of ML algorithms for the prediction of the abundance or presence/absence of HAB forming microalgae, include both biotic and abiotic variables. Abiotic variables used so far include site, season, calendar day, distance from the coast, meteorological variables (air temperature, wind speed and direction, cloud cover), physical variables of seawater (temperature, salinity, conductivity, turbidity, oxygen saturation, pH, Secchi depth, photosynthetic radiation, remote sensing reflectance) and nutrients in both dissolved and particulate forms [29,32,38,46]. Considering biotic variables, chl-*a* is often used [24,29,47], as well as the abundances of microcrustaceans, ciliates, tintinids, microheterotrophs, cladocerans or copepodes [29,36]. It seems that the selection of cause variables in the existing literature mainly aims to improving the prediction of HABs, although not necessarily exploring a cause-and-effect relationship from the ecological point of view.

In the present study, a screening of several ML algorithms and ensembles is performed aiming to detect the presence/absence of 18 harmful or potentially harmful microalgae at genus level, using existing data collected from six coastal areas in the Aegean Sea, Greece, representing different productivity levels. The overall performance of each algorithm is assessed considering all the 18 studied microalgae. The cause variables include physical variables (temperature and salinity) as well as nutrients, already established in previous studies [16,20] as drivers of algal growth. The objectives of the study are: (a) to identify the most efficient algorithms or ensembles based on the overall prediction of presence/absence of the 18 microalgae, (b) to quantify the effect of cause variables on microalgal growth and (c) to attempt possible grouping of the studied HAB forming microalgae, considering their response to the abiotic drivers. Since all cause variables are easily measured on a routine basis, the results of this study can form the basis of an operational system for the prediction of HABs in coastal waters, eliminating possible adverse effects on ecosystem and human health.

2. Materials and methods

2.1. Study area and sources of data

The data set used in the present work was set up by compilation of six sets collected from 42 sampling sites in three different areas of the Aegean Sea, Greece. Five out of six sets originate from coastal waters whereas one set (Rhodes offshore) is characteristic of pelagic waters (Figure 1). The sampling areas are: (a) Kalloni Gulf (Island of Lesbos): eight sampling stations K1-K8 [48] (b) Gulf of Gera (Island of Lesbos): Sampling sites G1-G8 [49] (c) the coastal area of the city of Mytilini: sampling sites M1 and M2 [50] (d) Saronikos Gulf (near Athens): sampling sites S1-S9 [51]

(e) offshore waters NW of the city of Rhodes R1-R5 [52] and (f) coastal waters of the city of Rhodes: ten sampling sites RH1-RH10 [51].

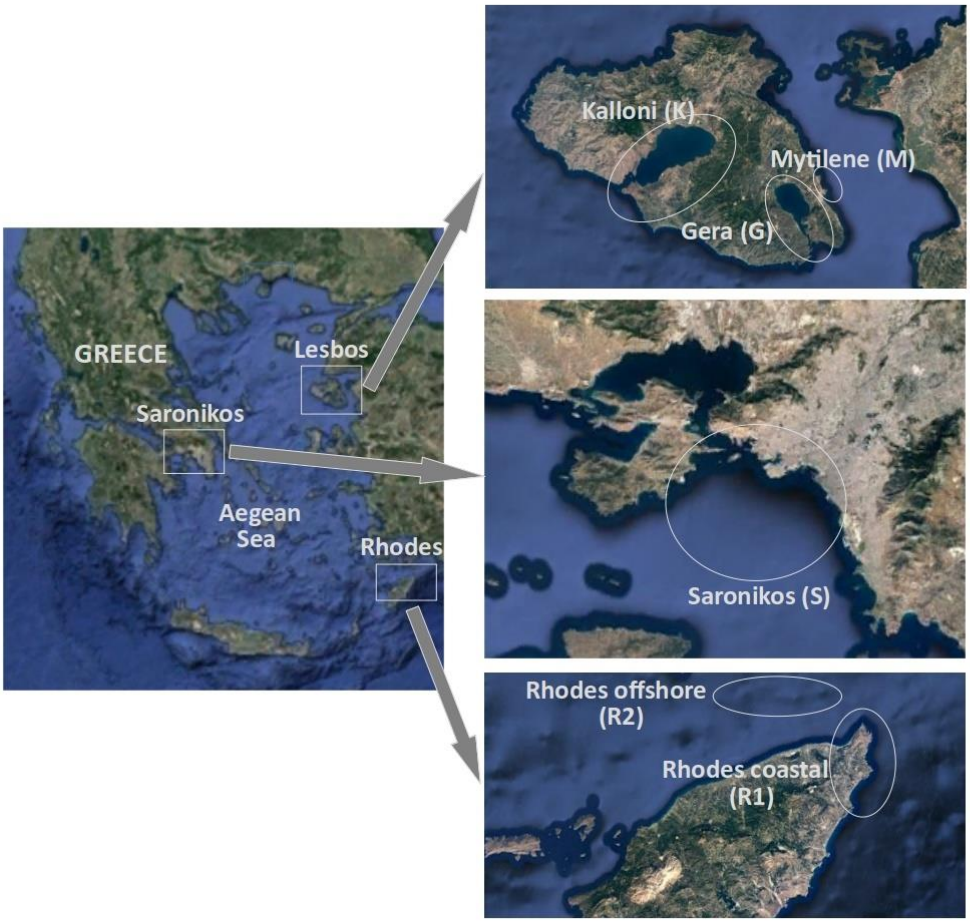

Figure 1. Map of the six sampling areas.

2.2. Variables used

The variables included in the dataset are: (a) temperature (T) and salinity (S) [20], also expressing seasonality and (b) nutrient concentrations, namely dissolved inorganic nitrogen (DIN), phosphates (PO_4^{3-}) and silicates (SiO_2), often recognized as drivers of algal growth and increased primary productivity in coastal waters [53]. These variables are used as inputs (cause variables) in the ML algorithms. The target variable trying to estimate is the presence or absence of 18 genera of harmful or potentially harmful algae estimated by the list of toxic microalgae of IOC-UNESCO [54], shown in Table 1. The number of occurrences of each genus in the 889 samples, the sites where found and summary statistics of the abiotic conditions under which each genus appeared, are shown in Table 1. Redfield' s ratio, i.e., nitrogen to phosphorus ratio (N:P), a characteristic proxy of the nutrient limiting primary productivity, is also shown in Table 1.

Table 1. The studied HAB forming microalgae, their occurrence and summary statistics (mean and range) of the corresponding abiotic variables related to each particular genus.

Genus	Occurrences	Sampling Areas	T (ºC)	S (psu)	DIN (μM)	PO_4^{3-} (μM)	SiO_2 (μM)	N:P
Alexandrium	25	K,R1	14.76	37.81	5.78	0.12	15.97	196.64
			(9.8-23.9)	(34.0-40.1)	(0.5-39.8)	(0.0-1.5)	(3.4-86.6)	(7.8-1217.5)
Amphidinium	71	K,S,R1,	17.04	38.26	4.78	0.13	14.34	232.41
		R2	(9.8-23.9)	(34.0-40.2)	(0.5-39.8)	(0.0-1.5)	(2.4-86.6)	(1.7-6020.0)
Cryptomonas	168	K	17.10	38.57	4.13	0.09	14.36	180.16
			(9.5-27.7)	(34.0-41.1)	(0.5-45.2)	(0.0-1.6)	(1.7-94.0)	(6.81-6020.0)
Dictyocha	48	K	13.93	38.29	4.46	0.10	16.22	102.69
			(9.4-23.3)	(34.8-40.7)	(0.6-33.2)	(0.0-1.3)	(4.3-70.9)	(7.4-692.2)
Dinophysis	228	K,G,M,	18.27	38.29	2.70	0.16	8.85	43.76
		S,R1,R2	(9.7-28.2)	(28.3-40.6)	(0.1-38.0)	(0.0-1.8)	(0.4-80.7)	(2.8-1712.0)
Goniodoma	21	G,M,	19.20	38.35	2.84	0.45	9.20	25.41
		S,R1,R2	(13.4-27.4)	(37.3-39.1)	(0.4-11.7)	(0.0-6.0)	(0.8-15.1)	(1.5-116.7)
Gonyaulax	37	K,G,M,	18.41	38.82	1.66	0.19	7.19	30.05
		S,R1,R2	(14.37-22.11)	(36.4-40.6)	(0.3-4.1)	(0.0-0.85)	(2.1-20.9)	(2.7-379.0)
Gymnodinium	279	K,M,	19.58	38.83	1.76	0.07	9.66	68.13
		S,R1,R2	(9.6-27.7)	(28.0-40.2)	(0.1-27.8)	(0.0-4.1)	(1.6-53.9)	(0.2-1660.0)
Gyrodinium	177	K,G,M,	18.23	38.51	2.78	0.16	8.63	70.01
		S,R1,R2	(9.7-27.7)	(34.7-40.3)	(0.4-45.2)	(0.0-1.8)	(1.5-94.0)	(1.7-1660.0)
Karenia	394	S,R1,R2	20.00	38.55	2.01	0.15	8.02	33.04
			(13.1-27.6)	(28.0-39.7)	(0.1-38.0)	(0.0-6.0)	(0.4-53.9)	(1.4-1643.0)
Karlodinium	133	K	17.20	38.68	3.51	0.07	12.64	180.74
			(9.9-27.7)	(34.8-40.7)	(0.5-33.2)	(0.0-1.3)	(1.7-70.9)	(6.8-6020.0)
Lingulodinium	23	K,R1	14.81	37.67	6.00	0.10	11.81	282.78
			(9.9-25.6)	(34.8-40.0)	(1.4-33.2)	(0.0-1.3)	(2.4-68.7)	(9.6-1660.0)
Navicula	165	G,M,	19.27	39.00	2.22	0.14	9.22	44.78
		S,R1,R2	(10.0-26.5)	(37.1-40.3)	(0.3-23.8)	(0.0-4.1)	(1.1-53.9)	(0.2-1643.0)
Peridinium	301	K,G,M,	18.92	38.62	2.60	0.11	9.87	71.58
		S,R1,R2	(9.6-27.7)	(28.0-40.3)	(0.1-39.8)	(0.0-2.3)	(1.2-86.6)	(1.7-1712.0)
Phaeocystis	71	S,R1	16.4	38.27	2.98	0.36	5.44	16.24
			(13.1-26.7)	(37.6-39.1)	(0.5-14.1)	(0.0-6.0)	(0.4-17.2)	(1.4-72.5)
Polykrikos	38	K,M,R1	14.88	38.7	3.77	0.06	9.56	346.81
			(10.0-25.6)	(34.9-40.4)	(0.6-22.8)	(0.0-0.9)	(1.1-70.9)	(11.9-6020.0)
Prorocentrum	777	K,G,M,	19.10	38.61	2.26	0.13	8.43	51.69
		S,R1,R2	(9.4-28.2)	(28.0-41.1)	(0.1-38.0)	(0.0-6.0)	(0.4-83.0)	(1.4-1660.0)
Pseudo-nitzschia	491	K,G,M,	17.39	38.5	2.89	0.14	8.92	78.51
		S,R1,R2	(9.4-27.7)	(28.3-41.1)	(0.4-45.2)	(0.0-6.0)	(0.4-94.0)	(1.4-1712.0)

2.3. Machine learning algorithms

A variety of ML algorithms was used to classify the presence or absence of the 18 genera of HAB forming microalgae (Table 2). The selected classifiers cover all basic supervised ML categories i.e., rules, trees, lazy, functions and bayes, whereas classic ensemble methods, already used in marine ecology [3], were also applied. In order to optimize the performance of the basic classifiers, different values of the crucial (hyper) parameters of each algorithm were tested during training. Thus, the training included at least four different versions of each basic algorithm (Table 2, last column) eliminating the relation of the parameters' selection to the final algorithm performance. Random Forest (RF) was considered as an ensemble, since it uses a set of decision trees to provide classification. For some ensembles (e.g., Voting), which use techniques trained by the combination of the results of basic classifiers, the number of combined classifiers was limited to three in order to avoid ties. The selected three classifiers were the three best performing basic classifiers, as previously defined.

2.4. Algorithm evaluation and selection of abiotic parameters

The efficiency of algorithms was evaluated using the 10-fold cross validation (10-fold CV) procedure [55]. The RWeka interface written in R [56], of Weka ML techniques [57], was used to run and test the algorithms. The number of correctly classified instances of either the presence or absence of each genus as labels of the target class over the total number of water samples (i.e., accuracy), was used to determine the performance effectiveness of each algorithm. Additionally, some other measures of predictive performance were used in order to better evaluate the algorithms' predictions [58]. More precisely, sensitivity (or recall) expressing algorithm completeness, is the fraction of the correct genus presences over the total number of predicted presences (true presences plus false absences) in the total samples. A similar measure for the prediction of absences was used as a second measure of completeness, since the conditions related to the absence of a genus in a sample are also considered crucial [59]. This measure is specificity which is the same fraction calculated on absences (number of correct predicted absences over the sum of true absences plus false presences). Moreover, precision was used, expressing the power of algorithm's correctness as it measures in how many instances that the algorithm predicted as "genus present", the genus was actually present. The classic measure of kappa statistic that represents the degree of accuracy and reliability in classification problems, was also included [60]. Kappa statistic ranges from -1 (total disagreement) through 0 (random classification) to 1 (perfect agreement). Finally, due to the existence of imbalanced data, i.e., low number of appearances for eight genera (i.e., *Alexandrium*, *Amphidinium*, *Dictyocha*, *Goniodoma*, *Gonyaulax*, *Lingulodinium*, *Phaeocystis* and *Polykrikos*) in the dataset (see Table 1), one more specialized evaluation measure was used; the discriminant power is a measure that combines sensitivity and specificity and evaluates how well an algorithm distinguishes the presences and absences of a rare genus in case of imbalanced data [59]. The thresholds for this measure are: "poor" for values less than 1, "limited" for values between 1 and 2, "fair" for values between 2 and 3 and "good" for values higher than 3 [61].

Table 2. ML algorithms used in the current study: category, abbreviation, short description and hyper parameter values.

Category	Abbreviation	Algorithm Description	Hyper parameter values
Rules	Jrip	Repeated incremental pruning to produce error reduction (RIPPER)	Batch size = 50, 100 Min total weight = 2, 5
	Part	PART decision list	Batch size = 50, 100 Min number of instances per rule = 2, 5
Trees	J48	C4 pruned decision tree	Batch size = 50, 100 Min number of instances per leaf = 2, 5 Number of folds = 3, 5
	Rep	Decision tree using reduced error pruning with backfiting	Batch size = 50, 100 Min number of instances per leaf = 2, 5
Lazy	IBk	The k-nearest neighbors using Euclidean distance	Number of neighbors = 1, 5, 10, 20
	KStar	Nearest neighbor with entropic distance	Global blend = 5, 10, 20, 50
Functions	Log	Multinomial logistic regression	
	MLP	Multilayer Perceptron using backpropagation	Number of neurons = 2, 5, 7, 10
Bayes	NB	Naïve Bayes using estimator classes	Batch size = 50, 100
	BN	Bayes network	Batch size = 50, 100
Ensembles	RF	Forest of random decision trees	Batch size = 50, 100 Number of iterations = 20, 50, 100
	Bagging	Bagging classifiers to reduce variance	Classifiers: The best basic one Number of iterations = 10
	Boosting	Boosting classifiers using Adaboost M1 method	Classifier: The best basic one
	CVR	Classification using regression methods	Classifier: M5 model tree
	RC	Randomizabling classifiers (Random Committee)	Classifier: Random tree
	Stacking	Combining classifiers using stacking method	Classifiers: The 3 best basic ones Number of folds = 10
	Voting	Combining classifiers using votes	Classifiers: The 3 best basic ones Combination rule: Average of Probabilities

The performance of an algorithm can be crucially affected by the variables included in the input dataset [24]. In order to select the optimal subset of the input abiotic variables offering the higher predictive performance to each trained algorithm, an exhaustive search was implemented. All possible subsets of the five available abiotic variables containing 5 (all input variables), 4 out of 5 (5 subsets), 3 out of 5 (10 subsets) and 2 out of 5 (10 subsets) variables were used to find the best combination.

2.5. Estimation of the importance of the cause variables

Some ML algorithms estimate directly the importance of the input variables (e.g., most rules or trees). Especially for tree algorithms, applying the split-criterion in each node, the relative importance of each input variable can be easily determined. The most common method to measure a node's effectiveness in trees is the Gini Impurity which measures the probability of misclassification for a new instance (i.e., sample) by its specific tree node. A related measure for RFs is the Mean Decrease of Impurity (MDI) which weighs the impact of each input variable to the final prediction (presence of a genus) by measuring the effectiveness of each variable at reducing the uncertainty during tree induction [62]. In this study the MDI measure was used to assess the importance of each abiotic variable to the presence or absence of the potentially harmful algae genera. Furthermore, this measure was also used for grouping genera affected by similar abiotic conditions. Hierarchical clustering was applied using Euclidean distance and Ward's minimum variance method.

3. Results

3.1. Performance of basic classifiers and ensembles

The performance of the ML algorithms measured by the accuracy of correctly classified presences/absences in the 889 seawater samples is presented in Table 3 while Table 4 contains the best performing algorithm along with the optimal values of hyper parameters and the optimal subset of the abiotic variables. The predictive performance for all genera is above 70%, being over 90% for 10 out of 18 genera. The higher percentage of correctly classified instances was for *Lingulodinium* (98.1%) and the lowest for *Peridinium* (72.2%). The performance of the various algorithms did not vary significantly within each genus, but mainly among genera. Indeed, the range of predictive performance among different algorithms for the same genus ranged from 32.2% for *Gymnodinium* to 1.4% for *Goniodoma* (9.1% the mean difference for all genera), while the corresponding range among genera for the same algorithm ranged from 44.2% for Naïve Bayes to 25.8% for RF technique (30.6% mean difference for all algorithms).

Among the basic algorithms, Jrip showed the best performance with mean predictive percentage for all genera equal to 87.7%. Part rule algorithm, both trees, KStar, MLP and BN also showed high performance exceeding 86%. Although IBK and MLP are the best performing algorithms regarding 6, and 3 genera respectively, they achieve lower overall performance compared to other basic algorithms. Finally, the classical statistical procedure of Logistic function and NB give predictions of rather low quality.

As expected, the ensembles generally exceed the performance of single classifiers, with four ensembles (RF, Bagging, Boosting, and Voting) achieving higher performances than the best basic, when considering means for all genera. The most efficient ensemble is RF with 89.2% mean predictive performance (1.5% higher than Jrip), providing the highest prediction for 9 genera (*Amphidinium*, *Dictyocha*, *Gonyaulax*, *Gymnodinium*, *Gyrodinium*, *Karenia*, *Karlodinium*, *Peridinium* and *Pseudo-nitzschia*). The second-best performing ensemble is voting with almost 0.7% lower mean predictive percentage (88.5%) compared to RF, offering the most successful prediction for *Phaeocystis*. Both Bagging and Boosting also performed sufficiently well, the former achieving

Table 3. Predictive performance (best performing single and ensemble in bold) of different ML techniques in terms of accuracy in HA genera presence/absence prediction as evaluated by 10-CV .

	Rules		Trees		Lazy		Functions		Bayes		Ensembles						
							Logisti				RF					Stackin	
HA genus	Jrip	Part	J48	Rep	IBK	KStar	c	MLP	NB	BN		Bagging	Boosting	CVR	RC	g	Voting
Alexandrium	0.970	0.973	0.972	0.972	0.976	0.968	0.971	0.972	0.957	0.965	0.973	0.972	0.968	0.972	0.972	0.972	0.972
Amphidinium	0.918	0.919	0.916	0.917	0.916	0.917	0.919	0.916	0.907	0.910	0.925	0.921	0.917	0.919	0.922	0.920	0.917
Cryptomonas	0.925	0.885	0.899	0.873	0.749	0.907	0.858	0.894	0.838	0.910	0.936	0.936	0.942	0.882	0.904	0.811	0.940
Dictyocha	0.954	0.943	0.947	0.943	0.955	0.953	0.943	0.943	0.936	0.948	0.965	0.946	0.946	0.940	0.943	0.946	0.937
Dinophysis	0.759	0.771	0.756	0.748	0.741	0.727	0.727	0.744	0.731	0.780	0.769	0.760	0.780	0.767	0.736	0.744	0.770
Goniodoma	0.976	0.976	0.976	0.976	0.979	0.970	0.975	0.975	0.965	0.976	0.976	0.978	0.979	0.977	0.970	0.976	0.976
Gonyaulax	0.957	0.955	0.958	0.958	0.955	0.954	0.956	0.958	0.926	0.926	0.961	0.960	0.961	0.958	0.953	0.958	0.958
Gymnodinium	0.772	0.780	0.793	0.772	0.692	0.803	0.680	0.742	0.524	0.753	0.846	0.811	0.773	0.812	0.802	0.686	0.827
Gyrodinium	0.793	0.802	0.802	0.800	0.729	0.801	0.800	0.798	0.775	0.799	0.827	0.812	0.790	0.799	0.790	0.801	0.808
Karenia	0.813	0.762	0.757	0.765	0.779	0.791	0.588	0.729	0.650	0.666	0.848	0.791	0.794	0.772	0.799	0.683	0.805
Karlodinium	0.901	0.879	0.880	0.889	0.748	0.898	0.857	0.884	0.852	0.916	0.917	0.915	0.916	0.899	0.911	0.850	0.916
Lingulodinium	0.980	0.981	0.981	0.978	0.974	0.972	0.973	0.974	0.966	0.974	0.980	0.976	0.975	0.974	0.981	0.980	0.981
Navicula	0.812	0.818	0.814	0.805	0.803	0.792	0.814	0.832	0.791	0.814	0.822	0.814	0.800	0.838	0.799	0.814	0.814
Peridinium	0.714	0.706	0.687	0.685	0.586	0.679	0.658	0.674	0.654	0.650	0.722	0.719	0.717	0.708	0.690	0.661	0.717
Phaeocystis	0.947	0.944	0.945	0.952	0.951	0.934	0.918	0.961	0.916	0.951	0.958	0.961	0.956	0.947	0.954	0.920	0.961
Polykrikos	0.957	0.962	0.961	0.961	0.964	0.955	0.958	0.958	0.945	0.958	0.962	0.961	0.958	0.962	0.958	0.960	0.963
Prorocentrum	0.876	0.875	0.874	0.875	0.876	0.880	0.874	0.874	0.862	0.874	0.877	0.877	0.872	0.875	0.854	0.874	0.874
Pseudo-nitzschia	0.757	0.756	0.761	0.754	0.771	0.765	0.665	0.699	0.565	0.737	0.791	0.768	0.771	0.762	0.750	0.653	0.786
Mean	0.877	0.872	0.871	0.868	0.841	0.870	0.841	0.863	0.820	0.862	0.892	0.882	0.879	0.876	0.872	0.845	0.885

an overall performance above 88.2% and the latter providing the best prediction for 4 genera. CVR, RC and Stacking show the same or lower performance compared with the basic algorithms and thus are considered as inefficient. For most genera (13 out of 18), the performance of the algorithms was optimal when all variables were included in the input data set. For *Dictyocha* and *Karenia* the best performance was achieved when removing DIN, whereas for the two Lazy algorithms (IBk and KStar) the prediction for *Alexandrium*, *Polykrikos* and *Prorocentrum* was optimal when T, S and PO_4^{3-} were only included in the input variables. For RF in particular, the use of all five abiotic variables optimized the prediction of 7 genera (*Ampidinium*, *Gonyaulax*, *Gymnodinium*, *Gyrodinium*, *Karlodinium*, *Peridinium* and *Pseudo-nitzschia*). Therefore in the rest of the analysis aiming to propose a single algorithm for all genera and to assess the relative importance of the input variables for prediction, RF the best performing ensemble was applied on the initial input data set of all five abiotic variables.

3.2. Assessing the performance of RF

Since RF seems to be the overall best performing algorithm, more details based on RF classification results using the initial dataset (optimal parameter combination) are shown in Table 5. The mean precision is 0.723, meaning that in 72.3% of the cases in which RF predicted the presence of a genus, the genus is actually present. For some genera this measure is rather high, being 91.5% for *Karlodinium* and 88.1% for *Prorocentrum*, showing an upward trend for genera with high number of occurrences in the whole database. On the other hand, the rate of correct classifications of presences that is sensitivity is rather moderate since only 43.1% of the total genus presences is classified correctly by the algorithm. The predictions are even worse for genera present in less than 10% of the samples (rare species), as *Alexandrium*, *Goniodoma*, *Gonyaulax* or *Polykrikos*. The specificity of the RF classification (correctly predicted absences) is generally high with a mean of 92.1%. This finding implies that absences are more correctly classified than presences, especially for rare genera. On the other hand, the mean value of Kappa statistic is equal to 0.469 showing a rather moderate predictive performance of the RF algorithm. The value of Kappa statistic is lower for the rare genera and varies from 0.287 (fair agreement) for *Gonyaulax* to 0.723 (substantial agreement) for *Cryptomonas.* However, the discriminant power, a special measure for the predictive performance for rare genera, is greater than 2 for 6 rare genera, showing that the prediction by RF is “fair” in these cases. Considering each specific genus, the prediction is “good” being above 3 for *Lingulodinium* and “limited” being 1.632 for *Amphidinium*.

Table 4. Best perfomed algorithme along with the optimal parameter combination for each genera.

HA genus	Best Algorithm	Hyper Paremeters values	Best Parameter Combination
Alexandrium	IBk	Number of neighbors=5	T+S+PO_4^{3-}
Amphidinium	RF	Batch size=50 Number of iterations=50	T+S+DIN+ PO_4^{3-}+SiO_2
Cryptomonas	Boosting	Classifier: Jrip (Batch size=100 Min total weight=5)	T+S+DIN+ PO_4^{3-}+SiO_2

Continued on next page

HA genus	Best Algorithm	Hyper Paremeters values	Best Parameter Combination
Dictyocha	RF	Batch size = 100 Number of iterations = 50	T + S + PO4 + SiO_2
Dinophysis	BN/ Boosting	Batch size = 100/ Classifier: IBk (Number of neighbors = 10)	T + S + DIN + PO_4^{3-} + SiO_2
Goniodoma	IBk	Number of neighbors = 10	T + S + DIN + PO_4^{3-} + SiO_2
Gonyaulax	RF	Batch size = 100 Number of iterations = 100	T + S + DIN + PO_4^{3-} + SiO_2
Gymnodinium	RF	Batch size = 50 Number of iterations = 50	T + S + DIN + PO_4^{3-} + SiO_2
Gyrodinium	RF	Batch size = 50 Number of iterations = 100	T + S + DIN + PO_4^{3-} +SiO_2
Karenia	RF	Batch size = 50 Number of iterations = 100	T + S + PO_4^{3-} + SiO_2
Karlodinium	RF	Batch size = 50 Number of iterations = 100	T + S + DIN+ PO_4^{3-} +SiO_2
Lingulodinium	Part/ RC/ Voting	Batch size = 50, Min number of instances per rule =2/ Classifier: Random tree/ Classifiers: Jrip (Batch size = 100, Min total weight = 5), Part (Batch size = 100, Min number of instances per rule = 2), J48 (Batch size = 100, Min number of instances per leaf = 5)	T + S + DIN + PO_4^{3-} +SiO_2
Navicula	CVR	Classifier: M5 model tree	T + S + DIN + PO_4^{3-} + SiO_2
Peridinium	RF	Batch size = 50 Number of iterations = 100	T + S + DIN + PO_4^{3-} + SiO_2
Phaeocystis	MLP/Bagging/Voting	Number of neurons = 7/ Classifier:MLP/ Classifiers: MLP, Rep (Batch size = 50, Min number of instances per leaf = 2), BN (Batch size = 100)	T + S + DIN + PO_4^{3-} + SiO_2
Polykrikos	IBk	Number of neighbors = 10	T + S + PO_4^{3-}
Prorocentrum	KStar	Global blend = 50	T + S + PO_4^{3-}
Pseudo-nitzschia	RF	Batch size = 50 Number of iterations = 100	T + S + DIN + PO_4^{3-} + SiO_2

Table 5. Predictive details of RF (the best performing algorithm) for each genus. The discriminant power refers only to genera with imbalanced data (rare genera).

Algae genus	N# presences	Precision	Sensitivity	Specificity	Kappa statistic	Discriminant power
Alexandrium	25	0.556	0.200	0.995	0.312	2.154
Amphidinium	71	0.567	0.239	0.984	0.386	1.632
Cryptomonas	168	0.878	0.774	0.974	0.723	
Dictyocha	48	0.615	0.333	0.988	0.503	2.049
Dinophysis	228	0.647	0.407	0.912	0.500	
Goniodoma	21	0.667	0.095	0.999	0.324	2.565
Gonyaulax	37	0.667	0.108	0.998	0.287	2.261
Gymnodinium	279	0.814	0.659	0.931	0.655	
Gyrodinium	177	0.682	0.353	0.970	0.509	
Karenia	394	0.758	0.787	0.800	0.604	
Karlodinium	133	0.915	0.489	0.992	0.595	
Lingulodinium	23	0.833	0.217	0.999	0.291	3.100
Navicula	165	0.550	0.200	0.963	0.480	
Peridinium	301	0.642	0.405	0.884	0.518	
Phaeocystis	71	0.867	0.549	0.993	0.633	2.840
Polykrikos	38	0.667	0.158	0.996	0.349	2.119
Prorocentrum	777	0.881	0.985	0.429	0.375	
Pseudo-nitzschia	490	0.811	0.808	0.769	0.668	
Mean values		0.723	0.431	0.921	0.469	2.340

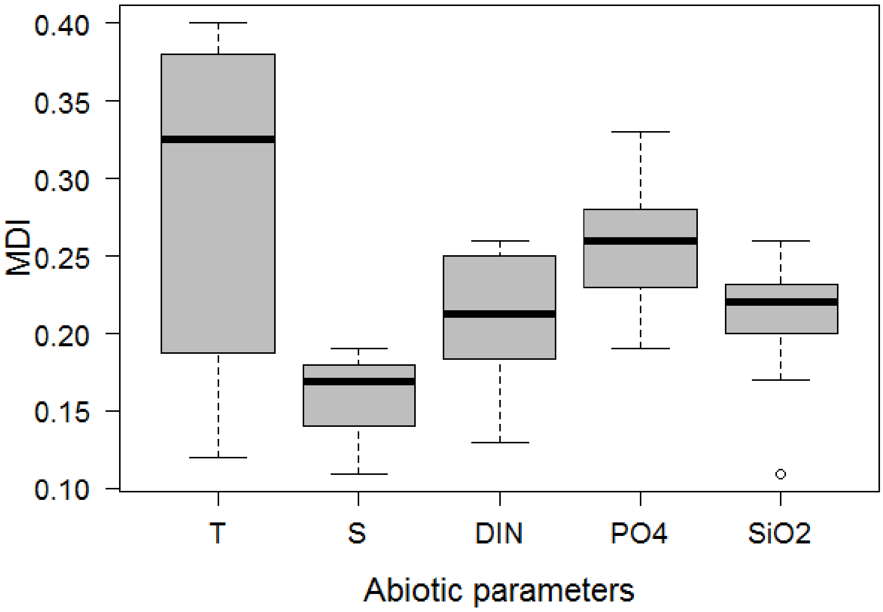

Figure 2. Box-and-whisker plots of MDI, measuring the relative importance of the various abiotic variables on the construction of RF trees.

3.3. Weighing the effect of abiotic variables

According to our findings, RF seems to be the most efficient algorithm among single classifiers and ensembles. Since RF is also showing a general immediacy in weighing the effects of input variables based on the composition of the constructed trees, it is selected for providing further insights on the role of the abiotic drivers for the presence/absence of potentially harmful microalgae.

The relative importance of each abiotic variable in the trees constructed for each genus by the RF algorithm, is assessed with the MDI measure (Figure 2). Temperature (T) has the most powerful effect with an MDI median of 0.325, whereas its importance varies among genera (minimum of 0.12 for *Pseudo-nitzschia* and maximum of 0.40 for *Prorocentrum)*. On the other hand, salinity (S) seems to have the lowest effect, with an MDI median equal to 0.169 (minimum of 0.11 for *Amfidinium* and maximum of 0.19 for *Lingulodinium)*. $PO_4^{3-}$is the most important variable among nutrients, with a median of 0.26, against 0.21 and 0.22 for DIN and SiO_2, respectively. Moreover, the importance of nutrients did not vary considerably among genera (standard deviation equal to 0.39, 0.39 and 0.34 for DIN, $PO_4^{3-}$and SiO_2, respectively), implying the similar effect of each nutrient on all species considered.

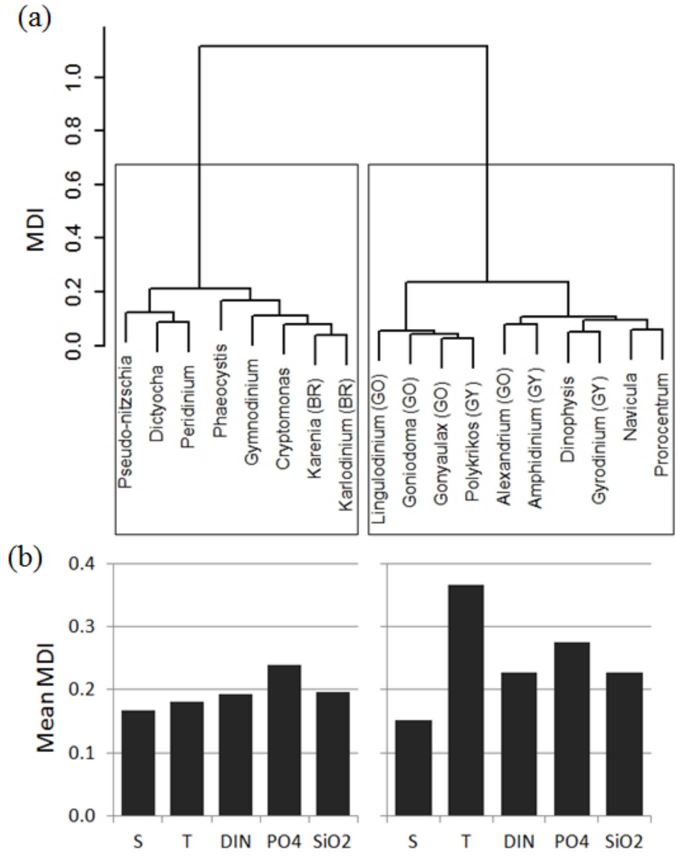

Figure 3. (a) Hierarchical clustering tree of the genera based on the relative importance of the abiotic variables based on MDI (Mean Decrease of Impurity) (b) The relative importance of the abiotic variables for each cluster. Brachidiniales (BR), Gonyaulacales (GO) and Gymnodiniales (GY) orders.

3.4. Clustering of genera according to the similarity of the effects of abiotic variables

Possible similarity in the importance of the effects of abiotic variables among the various genera in terms of MDI is assessed with hierarchical clustering (Figure 3). It seems that two groups are formed when considering the effect of the abiotic variables on trees' construction. The first cluster is formed off 8 genera, being *Pseudo-nitzschia*, *Dictyocha*, *Peridinium*, *Phaeoystis Gymnodinium*, *Cryptomonas*, and two genera belonging to the Brachidiniales (BR) order (*Karenia* and *Karlodinium*). The second cluster includes 10 genera, i.e., *Dinophysis*, *Navicula*, *Prorocentrum*, genera of the Gonyaulacales (GO) order (*Alexandrium*, *Goniodoma*, *Gonyaulax* and *Lingulodinium*) and of the Gymnodiniales (GY) order (*Amphidinium*, *Gyrodinium* and *Polykrikos*). The members of the first cluster are generally less affected by the abiotic variables compared to the members of the second cluster. Nutrients seem to have the main role for the occurrence of the members of the first cluster, compared to the physical variables (temperature and salinity). On the other hand, the members of the second cluster are mostly affected by temperature and nutrients, whereas salinity has the lowest effect.

4. Discussion

In this study, we assessed the overall performance of various ML techniques (both single and ensemble) for the prediction of the presence/absence of 18 harmful or potentially harmful marine microalgae at genus level. We further optimized the performance of the algorithms using different values of initial parameters using 10-fold CV. Jrip was the best performing basic classifier implying that single rules involving the five input variables can rather effectively model the response variable. The effectiveness of Jrip single rules has been also identified in previous studies [63,64]. The best ensemble method was RF achieving an overall mean accuracy of 89.2% and outperforming the rest algorithms by at least 1.5% in accuracy. This is in accordance with many previous studies that revealed the exceptional predictive power of RFs [65,33], including marine ecological studies [66,44]. A possible explanation for the successful performance of RFs is that they retain the benefits of the embedded decision trees, while they also combine tree results exploiting the great effectiveness of the majority of voting schema [67]. Thus, the recurring discrimination of samples based on the abiotic variables during the tree induction process crucially supports the detection of genera presences. The bootstrapping methods used for the construction of different trees protect RFs from overfitting issues while the majority of voting schema combines the predictions by taking into advantage the overlapped conditions that drive genus's occurrence in each tree [68]. Results on accuracy were remarkably high (89.2%), although the detailed analysis of the RF results revealed some predicting weaknesses. The sensitivity of the model was found rather moderate, implying that some real presences were not correctly identified by the algorithm. On the other hand, the specificity was above 90% indicating that absences were successfully determined. The precision of the predictions was found relatively high (72.3%), implying that the predicted presences are in a high percentage real presences. High specificity combined with moderate sensitivity and low kappa statistic is a weakness of RFs, usually arising when occurrences are rare [69,70]. Data bootstrapping is biased towards the major class (absences) leading RF technique to over-predict the major and under-predict the minor class [71]. This imbalance was observed in the present study since 8 genera (i.e., *Alexandrium*, *Amphidinium*, *Dictyocha*, *Goniodoma*, *Gonyaulax*, *Lingulodinium*, *Phaeocystis* and *Polykrikos*) were only present in less than 8% of the total samples. Discriminant power for these genera showed at least a "fair" prediction for RFs with the exception of *Amphidinium* for which prediction was characterized as "limited". This finding combined with the higher mean sensitivity for the other more common genera (being 58.7%), which improves as the balance between presences and absences increases (80.8% and 78.7% for the genera with the most presences *Pseudo-nitzschia* and *Karenia*, respectively) render RF results as satisfying. Therefore, special attention should be paid to the prediction of occurrences of rare genera using RFs or other ML techniques. This can be improved by using specific performance measures or by appropriate manipulations on the dataset, as oversampling or repeated random sampling [72]. The exhaustive search for the optimal subset of abiotic variables to be used for input in the ML algorithms, showed that the best performance was achieved when all five variables were used for most genera. As it was found in a similar study [24], a low number of input variables improves prediction, however it seems that in our case the number of five variables is already low and the removal of variables results to information loss. Aiming to exploit the advantages of RFs, we weighted the relative importance of the input variables (i.e., T, S, DIN, $\mathrm{PO_4^{3-}}$ and $\mathrm{SiO_2}$) using the initial dataset, in terms of MDI, for the presence/absence of the 18 harmful microalgal genera. Temperature was the most important driver of occurrence, although its power varied among genera. The key-role of temperature for the appearance and abundance of phytoplankton species has been recognized by various studies in Mediterranean Sea [20,73,48] and its effect is considered as mainly indirect, associated with seasonal changes and stratification which are identified as drivers of primary production in both ocean and coastal waters [74]. Due to the recent climate changes and the tendency of seawater temperature to increase, the effect of temperature is currently under thorough investigation [75]. It has been reported since a long time [76] that higher temperatures favor the growth of flagellates, whereas diatoms are well adapted at lower temperatures. According to Wyatt [77] "diatoms have seasonal standing-crop maxima in spring and autumn in middle latitudes while lower numbers occur during the intervening months". On the contrary blooms of dinoflagellates are more characteristic during summer time, especially in temperate and subtropical seas. The latter favors dinoflagellate dominance over diatoms during summer period. It is obvious that algal species succession is highly depended on temperature changes and this fact is clearly reflected in the results of the present work.

Although temperature is an important driver of HABs, salinity had the less significant role (lower mean MDI value) for the studied microalgae. There are two possible explanations for this lack of significance. Salinity acts on marine organisms through the osmotic pressure exercised on their cellular fluids. In the present work the range of salinity values was fairly limited (around 38 psu) and therefore it cannot be concluded whether this variable is not of importance to phytoplankton or the effects were not shown due to the narrow range of the salinity values used in this work.

Nutrients were also found to affect the presence of the studied microalgae, phosphates having a principal effect. The mean N:P values referring to specific genera are much higher than 16 with only one exception: the microalga *Phaeocystis* (with a corresponding N:P value 16:24). This value indicates that the conditions were phosphorus limited (Table 1). Nutrient concentrations in the surface waters in the Eastern Mediterranean are low due to mixing processes with nutrient poorer basin water and biological activity [78]. In addition to this shortcoming that limits phytoplankton abundance, the Eastern Mediterranean Sea seems to be phosphorus limited [79]. This is a rather peculiar characteristic, since the oceans, including the Atlantic, are nitrogen limited and the Atlantic Ocean supplies the Mediterranean with water masses through the Strait of Gibraltar. The interpretation of the results in the present work is also complicated because there is an additional problem regarding phosphates. There are some algal species that given a supply of phosphates, they can store it within their cells in the form of polyphosphate (volutin) granules [7]. This "luxury phosphorus" as it is known in the literature, can be consumed by the algae and support algal growth in the absence of an external supply source. This way any relationship between phosphorus concentration dissolved in the marine environment and algal standing stock is weakened due to the presence of stored phosphorus in some species.

According to the MDI values, expressing the effect of abiotic drivers for the appearance of the studied harmful microalgae, a possible grouping of those organisms was attempted. Two clusters were formed, the first of eight genera (*Pseudo-nitzschia*, *Dictyocha*, *Peridinium*, *Phaeoystis Gymnodinium*, *Cryptomonas*, *Karenia* and *Karlodinium*) and the second of ten genera (*Dinophysis*, *Navicula*, *Prorocentrum*, *Alexandrium*, *Goniodoma*, *Gonyaulax*, *Lingulodinium*, *Amphidinium*, *Gyrodinium* and *Polykrikos*). The first group is rather diverse, including members from several phyla, as Bacillariophyta (*Pseudo-nitzschia*), Ochrophyta (*Dictyocha*), Miozoa (*Peridinium*, *Gymnodinium*, *Karenia*, *Karlodinium*), Haptophyta (*Phaeocystis*) and Cryptophyta (*Cryptomonas*). The second group mainly includes members of the Miozoa phylum, except of *Navicula* belonging to Bacillariophyta. Therefore, common functional characteristics of genera belonging to the same phyla seem to play a role for their appearance [80], however this role is not so clear-cut. Members of both clusters are influenced by nutrients, and mainly by phosphorus, which is possibly related to the phosphorus limitation often observed in Eastern Mediterranean waters. For some of the studied genera, the role of phosphorus as a driver for presence of some genera can be found in the existing literature, as *Karlodinium* in Li et al. [81] and *Cryptomonas* in Gasol et al. (1993) [82]. Considering the physical drivers, salinity plays a minor role for both groups, although its role is important for specific microalgae as *Pseudo-nitzchia* [83,84] and *Karenia* [85]. Temperature is the main driver for the presence of the members of the second group, mainly involving Miozoa. Changes of temperature express seasonality, so related processes as stratification or factors as light availability [86] may be also important for the proliferation of the members of the second group.

Results from this study showed that it is possible to anticipate harmful algal blooms and design management practices to mitigate their effects on marine life and humans. This could be done by monitoring abiotic drivers that were identified in this work and issue appropriate warnings that may suggest some sort of action on behalf of appropriate management authorities. This study and the proposed methodology may form the basis for an effort to improve predictability of these occasionally devastating events.

5. Conclusions

In the present study various ML techniques were assessed for their efficiency to predict the appearance of 18 potentially harmful marine microalgae using data from 6 coastal sites in Eastern Mediterranean. RF algorithm using five abiotic drivers was the most efficient algorithm for prediction when considering overall the 18 studied microalgae. Moreover RF results were useful to enlighten the role of the various abiotic drivers in an ecological context. Although the overall performance of RF was satisfying in terms of predictability, a more exhaustive training of the algorithm with large number of samples is always desirable. Since the five abiotic variable are easily measured in a routine basis, the proposed methodology may form the basis of an operational system to be used for the prediction of HABs and therefore eliminate effects on marine life and human health.

Acknowledgments

This research is funded in the context of the project “Predicting harmful algae blooms using machine learning techniques” (MIS 5047187) under the call for proposals “Supporting researchers with an emphasis on new researchers” (EDULLL 103). The project is co-financed by Greece and the European Union (European Social Fund- ESF) by the Operational Programme Human Resources Development, Education and Lifelong Learning 2014–2020.

The authors wish to thank two anonymous reviewers for their insightful comments on a previous version of the manuscript.

Conflict of interest

There is no conflict of interests.

References

1. M. I. Jordan, T. M. Mitchell, Machine learning: Trends, perspectives and prospects, *Science,* **349** (2015), 255–260.
2. E. Alpaydin, *Introduction to machine learning,* 2nd Ed., The MIT Press, Cambridge, (2010).
3. C. Crisci, B. Ghattas, G. Perera, A review of supervised machine learning algorithms and their applications to ecological data, *Ecol. Modell.,* **240** (2012), 113–122.
4. G. P. Harris, Phytoplankton ecology: structure, function and fluctuation, *Chapman and Hall, London*, 1986.
5. K. H. Mann, J. R. N. Lazier, *Dynamics of marine ecosystems: biological-physical interactions in the Oceans*, Blackwell Scientific Publications, Oxford, (1991).
6. R. de Wit, L. J. Stal, B. A. Lomstein, R. A. Herbert, H. Van Gemerden, P. Viaroli, et al., ROBUST: The ROle of BUffering capacities in STabilising coastal lagoon ecosystems, *Cont. Shelf Res.*, **21** (2001), 2021–2041.
7. G. E. Fogg, B. Thake, *Algal culture and phytoplankton ecology*, 3rd Ed, The University of Wisconsin Press, (1987).
8. I. Valiela, *Marine ecological processes*, Springer-Verlag, New York, (1984).
9. J. M. Zaldivar, F. S. Bacelar, S. Dueri, D. Marinov, P. Viaroli, E. Hernández-García, Modeling approach to regime shifts of primary production in shallow coastal ecosystems, *Ecol. Modell.*, **220** (2009), 3100–3110.
10. G. M. Hallegraeff, Harmful algal blooms: a global overview, in *Manual on Harmful Marine Microalgae* (eds. G.M. Hallegraeff, D.M. Anderson and A.D. Cembella), UNESCO Publishing, (2003), 25–50.
11. P. Hoagland, S. Scatasta, The economic effects of harmful algal blooms, In *Ecology of Harmful Algae* (eds. Graneli E. and Turner J. T.), Springer-Verlag: Berlin, (2006), 391–401.

12. S. E. Shumway, A review of the effects of algal blooms on shellfish and aquaculture, *J. World Aquac. Soc.*, **21** (1990), 65–104.
13. D. Kitsiou, M. Karydis, Coastal marine eutrophication assessment: a review on data analysis, *Environ. Int.,* **37** (2011), 778–801.
14. L. Ignatiades, O. Gotsis-Skretas, A review on toxic and harmful algae in greek coastal waters (E. Mediterranean Sea), *Toxins, 2* (2010), 101–1037.
15. D. Kitsiou, H. Coccossis, M. Karydis, Multi-dimensional evaluation and ranking of coastal areas using GIS and multiple criteria choice methods, *Sci. Total Environ.*, **284** (2002) 1–17.
16. S. Spatharis, D. Mouillot, D. B. Danielidis, M. Karydis, T. Do Chi, G. Tsirtsis, Influence of terrestrial runoff on phytoplankton species richness-biomass relationships: a double stress hypothesis, *J. Exp. Mar. Biol. Ecol.,* **362** (2008), 55–62.
17. A. Menesguen, G. LacroiX, Modelling the marine eutrophication: a review, *Sci. Total Environ.,* **636** (2018), 339–354.
18. P. Jimeno-Saez, J. Senent-Aparicio, J. M. Cecilia, J. Perez-Sanchez, Using machine-learning algorithms for eutrophication modeling: case study of Mar Menor Lagoon (Spain*), Int. J. Environ. Res. Publ. Health.,* **17** (2020), 1189.
19. K. Rankinen, J. E. C. Bernal, M. Holmberg, K. Vuorio, Identifying multiple stressors that influence eutrophication in a Finnish agricultural river, *Sci. Total Environ.*, **658** (2019), 1278–1292.
20. A. Tamvakis, J. Miritzis, G. Tsirtsis, A. Spyropoulou, S. Spatharis, Effects of meteorological forcing on coastal eutrophication: modeling with model trees, *Estuar. Coast. Shelf Sci.,* **115** (2012), 210–217.
21. A. Catherine, M. Selma, D. Mouillot, M. Troussellier, C. Bernard, Patterns and multi-scale drivers of phytoplankton species richness in temperate peri-urban lakes, *Sci. Total Environ.,* **559** (2016), 74–83.
22. A. Tamvakis, V. Trygonis, J. Miritzis, G. Tsirtsis, S. Spatharis, Optimizing biodiversity prediction from abiotic parameters, *Environ. Model. Softw.,* **53** (2014), 112–120.
23. T.-H. Tran, N.-D. Hoang, Estimation of algal colonization growth on mortar surface using a hybridization of machine learning and metaheuristic optimization, *Sadhana,* **42** (2017), 929–939.
24. P. Yu, R. Gao, D. Zhang, Z.-P. Liu, Predicting coastal algal blooms with environmental factors by machine learning methods, *Ecol. Indic.,* **123** (2021), 107334.
25. M. Karydis, D. Kitsiou, Marine eutrophication: a global perspective, CRC Press (2020).
26. S. B. Watson, C. Miller, G. Arhonditsis, G. L. Boyer, W. Carmichael, M. N. Charlton, et al., The re-eutrophication of Lake Erie: harmful algal blooms and hypoxia, *Harmful Algae,* **56** (2016), 44–66.
27. T. Okaichi, Red tides, *Terra Scientific Publishing Company, Tokyo, Japan,* 2004.
28. N. Mellios, S. J. Moe, C. Laspidou, Machine Learning Approaches for Predicting Health Risk of Cyanobacterial Blooms in Northern European Lakes, *Water*, **12** (2020), 1191.

29. P. R. Hill, A. Kumar, M. Temimi, D. R. Bull, HABNet: Machine learning, remote sensing-based detection of harmful algal blooms, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **13** (2020), 3229–3239.
30. J. Derot, H. Yajima, S. Jacquet, Advances in forecasting harmful algal blooms using machine learning models: a case study with *Planktothrix rubescens* in Lake Geneva, *Harmful Algae*, **99** (2020), 101906.
31. L. Velo-Suarez, J. C. Gutierrez-Estrada, Artificial neural network approaches to one-step weekly prediction of *Dinophysis acuminata* blooms in Huelva (Western Andalucía, Spain), *Harmful Algae*, **6** (2007), 361–371.
32. M. Bourel, C. Crisci, A. Martinez, Consensus methods based on machine learning techniques for marine phytoplankton presence-absence prediction, *J. Mar. Syst.*, **42** (2017), 46–54.
33. D. R. Cutler, T. C. Edwards Jr, K. H Beard, A. Cutler, K. T. Hess, J. Gibson, et al., Random forests for classification in ecology, *Ecology*, **88** (2007), 2783–2792.
34. S. Lek, J. F. Guegan, Artificial neural networks as a tool in ecological modeling, an introduction, *Ecol. Modell.*, **120** (1999), 65–73.
35. A. Verikas, A. Gelzinis, M. Bacauskiene, I. Olenina, E. Vaiciukynas, An Integrated Approach to Analysis of Phytoplankton Images, *IEEE J. Ocean. Eng.*, **40** (2015), 315–326.
36. F. Recknagel, ANNA-Artificial neural network model for predicting species abundance and succession of blue-green algae, *Hydrobiologia*, **349** (1997), 47–57.
37. C. Guallar, M. Delgado, J. Diogene, M. Fernandez-Tejedor, Artificial neural network approach to population dynamics of harmful algal blooms in Alfacs Bay (NW Mediterranean): Case studies of Karlodinium and Pseudo-nitzschia, *Ecol. Model.*, **338** (2016), 37–50.
38. H. M. Oh, C. Y. Ahn, J. W. Lee, T. S. Chon, K. H. Choi, Y. S. Park, Community pattering and identification of predominant factors in algal bloom in Daechung Reservoir (Korea) using artificial neural networks, *Ecol. Modell.*, **203** (2007), 109–118.
39. J. L Degling, C. Jin, A. Wong, Investigating the automatic classification of algae using the spectral and morphological characteristics via deep residual learning, in *Image Analysis and Recognition. International Conference on Image Analysis and Recognition* (eds. F. Karray, A. Campilho, A. Yu), Lecture Notes in Computer Science (11663), Springer, (2019), 269–280
40. X. Li, R. Liao, J. Zhou, P. T. Leung, M. Yan, H. Ma, Classification of morphologically similar algae and cyanobacteria using Mueller matrix imaging and convolutional neural networks, *Appl. Optics*, **56** (2017), 6520–6530.
41. A. El-habashi, I. Ioannou, M. C. Tomlinson, R. P. Stumpf, S. Ahmed, Satellite retrievals of Karenia brevis harmful algae blooms in the west Florida shelf using neural networks and comparison with other techniques, *Remote Sens.*, **8** (2016), 377.
42. J. M. T. Palenzuela, L. G. Vilas, F. M. B. Aláez, Y. Pazos, Potential Application of the New Sentinel Satellites for Monitoring of Harmful Algal Blooms in the Galician Aquaculture, *Thalassas*, **36** (2020), 85–93.

43. S. Hu, H. Liu, W. Zhao, T. Shi, Z. Hu, Q. Li, et al., Comparison of machine learning techniques in inferring phytoplankton size classes, *Remote sens.*, **10** (2018), 191
44. B. Bejaoui, Z. Armi, E. Ottaviani, E. Barelli, E. Gargouri-Ellouz, R. Cherif, et al., Random forest model and TRIX used in combination to assess and diagnose the trophic status of Bizerte Lagoon, southern Mediterranean, *Ecol. Indic.*, **71** (2016), 293–301.
45. H. Yajima, J. Derot, Application of the Random Forest model for chlorophyll-a forecast in fresh and brackish water bodies in Japan, using multivariate long-term databases, *Hydroinformatics*, **20** (2018), 206–220.
46. G. Martinez de la Escalera, C. Kruk, A. M. Segura, L. Nogueira, I. Alcantara, C. Piccini, Dynamics of toxic genotypes of Microcystis aeruginosa complex (MAC) through a wide freshwater to marine environmental gradient, *Harmful Algae*, **62** (2017), 73–83.
47. E. Valbi, F. Ricci, S. Capellacci, S. Casabianca, M. Scardi, A. Penna, A model predicting the PSP toxic dinoflagellate Alexandrium minutum occurrence in the coastal waters of the NW Adriatic Sea, *Scientific Reports*, **9** (2019), 4166.
48. S. Spatharis, N. P. Dolapsakis, A. Economou-Amilli, G. Tsirtsis, D. B. Danielidis, Dynamics of potentially harmful microalgae in a confined Mediterranean Gulf-Assessing the risk of bloom formation, *Harmful Algae*, **8** (2009), 736–743.
49. G. Arhonditsis, G. Tsirtsis, M. Karydis, The effects of episodic rainfall events to the dynamics of coastal marine ecosystems: applications to a semi-enclosed gulf in the Meditteranean Sea, *J. Mar. Syst.*, **35** (2002), 183–205.
50. G. Tsirtsis, M. Karydis, Evaluation of phytoplankton community indices for detecting eutrophic trends in the marine environment, *Environ. Monit. Assess.*, **50** (1998), 255–269.
51. M. Karydis, Quantitative assessment of eutrophication: a scoring system for characterising water quality in coastal marine ecosystems, *Environ. Monit. Assess.*, **41** (1996), 233–246.
52. D. Kitsiou, M. Karydis, Categorical mapping of marine eutrophication based on ecological indices, *Sci. Total Environ.*, **255** (2000), 113–127.
53. S. Spatharis, G. Tsirtsis, D. Danielidis, T. Do Chi, D. Mouillot, Effects of pulsed nutrient inputs on phytoplankton assemblage structure and blooms in an enclosed coastal area, *Estuar. Coast. Shelf Sci.*, **73 (**2007), 807–815.
54. Ø. Moestrup, R. Akselmann-Cardella, C. Churro, S. Fraga, M. Hoppenrath, M. Iwataki, et al., *IOC-UNESCO Taxonomic Reference List of Harmful Micro Algae*, (2009), Available from: http://www.marinespecies.org/hab on 2021-04-24.
55. M. Stone, Cross-validation and multinomial prediction, *Biometrica*, **61** (1974), 509–515
56. K. Hornik, C. Buchta, A. Zeileis, Open-source machine learning: R Meets Weka. *Comput. Stat.*, **24** (2009), 225–232.
57. I. H. Witten, E. Frank, Data mining: practical machine learning tools and techniques, 2nd edition, *Morgan Kaufmann, San Francisco*, 2005.
58. B. Juda, H. S. Le, Precision-recall versus accuracy and the role of large data sets, *Proc. AAAI Conf. Artif. Intell.*, **33** (2019), 4039–4048.

59. M. Bekkar, H. K. Djemaa, T. A. Alitouche, Evaluation measures for models assessment over imbalanced data sets, *J. Inf. Secur. Appl.*, **3** (2013), 27–39.

60. J. Cohen, A coefficient of agreement for nominal scales, *Educ. Psychol. Meas.*, **20** (1960), 37–46.

61. R.O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, 2nd Edition, Wiley-Interscience, USA, (2000).

62. G. Louppe, L. Wehenkel, A. Sutera, P. Geurts, Understanding variable importances, in *Forest of randomized trees* (eds C.J.C Burges, L. Bottou, M. Welling, Z. Chahramani and K.Q. Weinberger), Advances in Neural Information Processing Systems, (2013) 431–439.

63. W. N. H. W. Mohamed, M. N. M. Salleh, A. H. Omar, A comparative study of Reduced Error Pruning method in decision tree algorithms, *IEEE Int. Conf. Control Syst. Comput. Eng.*, Penang, (2012), 392–397.

64. P. Jain, J. M. Garibaldi, J. D. Hirst, Supervised machine learning algorithms for protein structure classification, *Comput. Biol. Chem.*, **33** (2009), 216–223.

65. M. Belgiu, L. Dragut, Random forest in remote sensing: A review of applications and future directions, *ISPRS J. Photogramm. Remote Sens.*, **114** (2016), 24–31.

66. K. Miller, F. Huettmann, B. Norcross, M. Lorenz, Multivariate random forest models of estruarine-associated fish and invertebrate communities, *Mar. Ecol. Prog. Ser., **500*** (2014), 159–174.

67. Y. Qi, Random Forest for Bioinformatics, *In Ensemble Machine Learning. (*eds C. Zhang, Y. Ma) Springer, Boston, MA, (2012).

68. P. Yang, Y. Hwa Yang, B. B. Zhou, A. Y. Zomaya, A review of ensemble methods in bioinformatics, *Curr. Bioinform.*, **5** (2010), 296–308.

69. H. R. Sofaer, J. A. Hoeting, C. S. Jarnevich, The area under the precision-recall curve as a performance metric for rare binary events, *Methods Ecol. Evol.*, **10** (2018), 565–577.

70. Q. Gu, L. Zhu, Z. Cai, Evaluation measures of the classification performance of imbalanced data sets, in *Computation Intelligence and Intelligent Systems vol 51 (*eds Z.Cai, Z. Li, Z. Kang and Y. Liu), Springer, Berlin, Heidelberg, (2009).

71. C. Chen, A. Liaw, L. Breiman, *Using random forest to learn imbalanced data*, University of California, Berkeley, (2004).

72. M. Khalilia, S. Chakraborty, M. Popescu, Predicting disease risks from highly imblanced data using random forest, *BMC Medical Inform. Decis. Mak.*, **11** (2011), 51.

73. S. Moncheva, O. Gotsis-Skretas, K. Pagou, A. Krastev, Phytoplankton blooms in Black Sea and Mediterranean coastal ecosystems subjected to anthropogenic eutrophication: similarities and differences, *Estuar. Coast. Shelf Sci.*, **53** (2001), 28–295.

74. K. H. Mann, J. R. N. Lazier, 2005. Dynamics of Marine Ecosystems. *Blackwell Publishing Ltd.*

75. C. Marampouti, A. C. J. Buma, M. Karin de Boer, Mediterranean alien harmful algal blooms: origins and impacts, *Environ. Sci. Pollut. Res.*, **28** (2021), 3837–3851.

76. D. H. Cushing, J. J. Walsh, *The ecology of the seas*. Blackwell Scientific Publications, Oxford, (1976).
77. T. Wyatt, Plants and animals of the sea, in: *The ecology of the seas* (eds. D.H. Cushing and J.J. Walsh), Blackwell Scientific Publications, Oxford, (1976), 81–97.
78. UNEP, *State and pressures of the marine and coastal Mediterranean environment*, Environmental Assessment Series (No. 5), European Environment Agency, Copenhagen. (1999).
79. B. R. Berland, J. Bonin, S. Y. Maestrini, Azote ou phosphore? Considerations sur le paradoxe nutritionnel de la Mediterranee, *Oceanol. Acta*, **3** (1980), 135–142.
80. E. Litchman, P. de Tezanos Pinto, C. A. Klausmeier, M. K. Thomas, K. Yoshiyama, Linking traits to species diversity and community structure in phytoplankton. *In: L. Naselli-Flores, G. Rossetti (eds) Fifty years after the "Homage to Santa Rosalia": Old and new paradigms on biodiversity in aquatic ecosystems. Developments in Hydrobiology 213*, Springer, Dordrecht, (2010), 12–28.
81. J. Li, P. M. Glibert, Y. Gao, Temporal and spatial changes in Chesapeake Bay water quality and relationships to Prorocentrum minimum, Karlodinium veneficum and CyanoHAB events, 191-2008. *Harmful Algae*, **42**, (2015), 1–14.
82. J. M. Gasol, J. Garcia-Cantizano, R. Massan, R. Guerrero, C. Pedros-Alio, Physiological ecology of a metalimnetic Cryptomonas population: relationships to light, sulfide and nutrients, *J. Plankton Res.*, **15** (1993), 255–275.
83. S. M. Pednekar, S. S. Bates, V. Kerkar, S. G. P. Matondkar, Environmental factors affecting the distribution of Pseudo-nitzschia in two monsoonal estuaries of Western India and effects of salinity on growth of domoic acid production by *P. pungens*, *Estuaries Coasts*, **41** (2018), 1448–1462.
84. C. R. Anderson, M. R. P. Sapiano, M. B. K. Prasad, W. Long, P. J. Tango, C. W. Brown, et al., Predicting potentially toxigenic *Pseudo-nitzchia* blooms in the Chesapeake Bay, *J. Mar. Syst.*, **83** (2010), 127–140.
85. W. Feki-Sahnoun, H. Njah, A. Hamza, N.Barraj, M. Mahfoudi, A. Rebai, et al., Using general linear model, Bayesian Networks and Naïve Bayes classifier for prediction of *Karenia Selliformis* occurrences and blooms, *Ecol. Inform.*, **43** (2018), 12–23.
86. G. M. Grimaud, F. Mairet, A. Sciandra, O. Bernard, Modeling the temperature effect on the specific growth rate of phytoplankton: a review, *Rev. Environ. Sci. Biotechnol.*, **16** (2017), 625–645.