Comparative analysis of machine learning models for time-series forecasting of *Escherichia coli* contamination in Portuguese shellfish production areas

Filipe Ferraz, Susana Vinga and Alexandra M. Carvalho

***Abstract*—Shellfish farming and harvesting have experienced a surge in popularity in recent years. However, the presence of faecal bacteria can contaminate shellfish, posing a risk to human health. This can result in the reclassification of shellfish production areas or even prohibit harvesting, leading to significant economic losses. Therefore, it is crucial to establish effective strategies for predicting shellfish contamination by the bacteria *Escherichia coli* (*E. coli*). In this thesis, various univariate and multivariate time series forecasting models were investigated to address this problem. These models include the autoregressive integrated moving average (ARIMA), vector autoregressive (VAR), random forest, and artificial neural networks (ANN) like the feed forward (FFNN) and long short-term memory (LSTM) networks. The data used in this thesis consisted of measurements of both *E. coli* concentrations and meteorological variables provided by the Portuguese Institute of Sea and Atmosphere (IPMA) regarding three shellfish production areas. Overall, the autoregressive models achieved the lowest root mean squared error (RMSE) and good classification results across all experiments. Moreover, in general, the multivariate ANNs outperformed the univariate ones, with one multivariate FFNN obtaining a true positive rate (TPR) of 1 and a true negative rate (TNR) of 0.75 in one area. This work represents the initial steps in the search for candidate forecasting models to help the shellfish production sector in anticipating harvesting prohibitions and, hence, supporting management and regulation decisions.**

Index Terms*—Time Series, Forecasting, Shellfish Contamination, *E. coli

I. Introduction

OVER the past few years, shellfish farming and harvesting in aquaculture have experienced significant growth in both quantity and economic value, primarily due to the continuous rise in demand for fish protein among humans [1]. While livestock production demands large volumes of freshwater, marine aquaculture is the food-producing sector least dependent on its availability since freshwater is not used in the farming process of shellfish or previous stages of the production chain [2]. Shellfish farming also attenuates ocean eutrophication - an excessive richness of nutrients in a body of water, often caused by runoff from land, resulting in dense plant growth and animal death due to lack of oxygen. This is because shellfish are filter-feeding organisms relying on the plankton in the water column to grow [2; 3].

However, microbiological contamination and the proliferation of toxic phytoplankton can compromise water quality and, consequently, the sustainability of shellfish farming businesses by causing the prohibition of harvesting or reclassification of production areas into worse sanitary statuses. Furthermore, consuming contaminated shellfish, whether from faecal bacteria or biotoxins, poses a risk to human health. Faecal coliforms, including *Escherichia coli* (*E. coli*), are helpful indicators for evaluating the bacterial quality of shellfish harvesting areas [4].

In fact, in the European Union, the sanitary quality of these areas is determined by monitoring *E. coli* levels in the shellfish flesh [5]. Although most contaminants may naturally depurate and dissipate over time while shellfish are still living in the water, it is often not financially feasible to return them to the water once they have been harvested [6].

To ensure public health safety, the Portuguese Institute of Sea and Atmosphere (IPMA) regularly monitors shellfish, classifies the production areas and prohibits their harvest and commercialisation if the *E. coli* concentrations, expressed as most probable number (MPN) per 100 g, exceed the safety limits defined in EU Regulations [5] (refer to Table I). Although the current strategy is effective in consumer protection, it is reactive, thus responding only after shellfish are harvested, resulting in severe economic losses. Therefore, it is imperative to develop predictive strategies for shellfish contamination.

TABLE I: Regulatory limits for *E. coli* concentrations in shellfish [5].

Sanitary Status	*E. coli* Regulatory Limits	Observations
A	80% of the results $\leq$ 230 (MPN/100 g) and 100% of the results $\leq$ 700 (MPN/100 g)	Bivalves can be caught and marketed for direct human consumption
B	90% of the results $\leq$ 4600 (MPN/100 g) and 100% of the results $\leq$ 46000 (MPN/100 g)	Bivalves may be harvested and destined for purification, transposition or processing into an industrial unit
C	100% of the results $\leq$ 46000 (MPN/100 g)	Bivalves may be harvested and intended for prolonged transposition or transformation into an industrial unit only
Forbidden	Any result> 46000 (MPN/100 g)	The harvest of bivalves is not permitted

A. Related Work

Artificial intelligence can play a crucial role in predicting shellfish contamination. Thanks to its ability to analyse vast and varied datasets, it can identify patterns and trends that might otherwise go unnoticed. Therefore, based on environmental conditions, it can be used to assess the risk of *E. coli* contamination in different regions. This information can help authorities and local producers prioritise resources

and take preventive measures to reduce the consequences of contamination.

Several modelling approaches have been proposed to forecast environmental phenomena in aquatic ecosystems, as is the case of predicting chlorophyll a (chl-a) concentration, extensively studied for water quality purposes. Chen et al. (2015) [7] developed an ARIMA model to predict the daily chl-a concentrations with data from Taihu Lake in China. The ARIMA model outperformed a multivariate linear regression (MVLR) model in terms of predictive accuracy.

In addition, Bourel et al. (2021) [8] introduced and evaluated several machine learning methods to model imbalanced data for predicting faecal contamination in beach waters in Uruguay using meteorological variables. From the tested models, the stratified Random Forest presented the best performance, improving the true positive rate at 50% concerning the baseline.

Due to ANN's ability to model non-linear relationships in the data, its use to solve complex time series forecasting problems has been increasing. Thoe et al. (2012) [9] proposed Multiple Linear Regression (MLR) and ANN (specifically a three-layer FFNN) models to predict the next-day *E. coli* concentration on four beaches in Hong Kong: Big Wave Bay, Deep Water Bay, New Cafeteria and Silvermine Bay. The authors used seven hydro-environmental variables that strongly correlated with *E. coli* as input, from monitoring data between 2002 and 2006. Results showed that the models could track the dynamic changes in *E. coli* concentration.

Additionally, Cho et al. (2018) [10] applied an LSTM to predict the concentration of *chl-a* using daily measured water quality data from the Gongju observation station of the Geum River (South Korea). The authors conducted a comparison of their results with previous approaches in predicting *chl-a* concentration one and four days in advance. The model used in this study demonstrated superior performance. In another work, Lee and Lee (2018) [11] developed deep learning models to predict one-week-ahead *chl-a* concentration, a well-established indicator for algal activity, in four major rivers of South Korea. This was accomplished by utilizing water quality and quantity data. The structure of the models consisted of nine input variables, three hidden layers (the first with 32 nodes and the remaining two with 64 nodes), and one output layer, all of which were completely connected. The findings revealed that the LSTM model outperformed the other two models, following the trend line even when variations in *chl-a* were large. Regarding the continental Portuguese coast, Cruz et al. (2022) [12] developed multiple forecasting methods to predict mussel contamination by *diarrhetic shellfish poisoning* (DSP) toxins. The data used consisted of DSP toxin concentration, toxic phytoplankton cell counts, meteorological variables and remote sensing data such as SST and chl-a concentration. The results showed that the artificial neural network models (ANN) outperformed the VAR and ARIMA models. Surprisingly, the LSTM model, trained only on the biotoxin variable, outperformed the multivariate models by accurately predicting DSP concentration one-week-ahead.

Several studies have highlighted the relationship between *E. coli* concentrations and environmental variables, many of which are utilized in this work. Anacleto et al. (2013) [13] investigated how seasonal and environmental parameters influence the occurrence of *E. coli*, among other bacteria, in two clam species, water and sediment from the Tagus estuary in Portugal. The authors concluded that overall, the levels of *E. coli* were higher in samples collected in colder months, revealing higher dissolved oxygen, pH and rainfall, along with lower temperature and salinity. Other studies, such as the ones developed by Jang et al. (2022) [14] and Campos et al. (2013) [15], also present low temperature, salinity and solar radiation, as well as rainfall events, as key factors for the survival of faecal indicator organisms such as the *E. coli*. Furthermore, in the already mentioned work of Thoe et al. (2012) [9], the data analysis showed that *E. coli* strongly correlated with seven of the thirteen variables studied: rainfall, solar radiation, wind speed, tide level, salinity, water temperature and past *E. coli* concentration.

B. Proposed Approach

This study attempts to predict shellfish contamination by *E. coli* in Portugal, which has not been done before. Based on the positive outcomes of previous literature, this work utilizes ARIMA, VAR, Random Forest FFNN and LSTM models to forecast the concentration of *E. coli* one month in advance in three shellfish production areas along the continental Portuguese coast.

The work was developed in *Python* and the remaining of the document is structured as follows. Section II provides a theoretical overview of the concepts necessary to properly understand the conducted study. Section III describes the data used for the study, the steps taken to prepare them for the prediction models and the development of the models. Section IV shows and discusses the results obtained by the forecasting models. Finally, the conclusions and suggestions for future work are presented in Section V.

II. Theoretical Overview

A. Time Series

A time series (TS) is a set of observations, each being recorded at a specific time. A TS can be either discrete or continuous depending on whether the observations are recorded only at specific time intervals or continuously over some period of time [16].

TS can be univariate (UTS) if they contain only values of a single variable over time or multivariate (MTS) when m variables are measured over time [17]. A UTS can be represented as $\{X_t\}_{t\in\{1,\dots,T\}}$ and an MTS as a family $\{\boldsymbol{X}_t\}_{t\in\{1,\dots,T\}}$ containing multiple UTS, each denoted by the vector $\boldsymbol{X}_t = (X_{1t}, X_{2t}, \dots, X_{mt})$, where X_{it} is the *i*-th component variable at time *t* [18]. The key feature of an MTS is that its observations are not only dependent on component *i*, but also on time *t* [18]. This implies that both the correlation between observations of a single component variable X_{it} at different times and the interdependence between different component series X_{it} and $X_{jt'}$ must be taken into account. Even when *t* and *t'* are not the same, these series can be correlated if $i \neq j$ [18].

B. Time Series Models

ARIMA is a UTS model which combines an autoregressive model (AR) with a moving average model (MA). A TS $\{X_t\}_{t\in\{1,\dots,T\}}$ is said to be an AR(p) if it is a weighted linear sum of the past p values plus a random shock Z_t with zero mean and constant variance [17], while a MA(q) corresponds to a TS that is a weighted linear sum of the last q random shocks (with zero mean and constant variance) [17]. By adding together the terms of an AR(p) model and a MA(q) model, we obtained an autoregressive moving average (ARMA) process of order (p, q), denoted ARMA(p, q) [19], given by:

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t + \cdots + \theta_q Z_{t-q}. \quad (1)$$

In practice, most TS exhibit non-stationarity, making it unsuitable for applying AR, MA, or ARMA models directly. These models are only appropriate for data without trends and seasonality. However, a non-stationary TS can be transformed into a stationary one by differencing adjacent terms. If the original data series is differenced d times before fitting an ARMA(p, q) model, then the model for the original undifferenced series is said to be an ARIMA(p, d, q) model [17].

The models presented so far are only applicable to UTS. The VAR model extends the AR model to the multivariate context. A MTS $\{\boldsymbol{X}_t\}_{t\in\{1,\dots,T\}}$ follows a VAR model of order p, VAR(p), if

$$\boldsymbol{X}_t = \boldsymbol{\phi}_0 + \sum_{i=1}^{p} \boldsymbol{\phi}_i \boldsymbol{X}_{t-i} + \boldsymbol{a}_t, \quad (2)$$

where $\boldsymbol{\phi}_0$ is a m-dimensional constant vector, $\boldsymbol{\phi}_i$ is a $m \times m$ matrix for $i > 0$ and $\boldsymbol{a}_t$ is a sequence of independent and identically distributed random vectors with zero mean and a positive definite covariance matrix [20].

The Random Forest (RF) model is an ensemble tree-based learning algorithm. A tree-based model involves recursively partitioning a given dataset into two groups based on a particular criterion until a predetermined stopping condition is met [21]. However, decision trees are prone to overfitting, which means that the model follows the peculiarities of the training dataset too closely and performs poorly on a new dataset, not being able to generalize well. One way to solve this is by considering only a subset of the observations and building many individual trees, which is what the random forest model does. Finally, the model's output is determined by averaging predictions over many individual trees [21].

An Artificial Neural Network (ANN) is a model composed of many simple non-linear computing units called neurons. By organising the neurons in layers, we form a multilayer perceptron (MLP) or feed-forward neural network (FFNN). An FFNN comprises the first layer, which receives inputs from the environment; the last layer, which outputs the network's prediction; and the layers in between (hidden layers). Each neuron in each layer is connected to each neuron in the following layer, and each connection has an associated weight. The goal is to adjust the weights so that the neural network's output is as close as possible to the desired result, which happens during training through the back-propagation algorithm [22].

LSTM networks are a type of recurrent neural network (RNN) that can learn long-term dependencies. An LSTM cell consists of a hidden state, which represents the short-term memory component, and an internal cell state, which represents the long-term memory [23; 24; 25]. Each cell is equipped with a set of gating units that regulate the flow of information, namely the input, forget, and output gates. The input and forget gates work in tandem to decide the amount of previous information to preserve in the current cell state and the amount of current context to transmit to future time steps [24; 25].

C. Performance Measuring Methods

Evaluating the performance of the aforementioned models is essential. The mean squared error (MSE) is a loss function that calculates the average squared difference between estimated and true values, while the root mean squared error (RMSE) represents its square root, given by

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \quad (3)$$

where n represents the number of samples in the set, and y_i and $\hat{y}_i$ are the observed and estimated values for samples i, respectively.

Table I provides four possible categories for the predicted values of *E. coli*, allowing the problem at hand to be formulated as a multiclass classification problem instead of a regression one. In this thesis, only two classes rather than four were considered: class A, which covers the points less than or equal to 230 (MPN/100 g), and class B, covering the remaining points (concentrations above 230 MPN/100 g). The classification metrics used in this thesis were the true positive rate (TRP) and the true negative rate (TNR) given by 4 and 5. Lastly, the amount of correctly labelled points of transition between classes was also used as a metric to evaluate the models.

$$TPR = \frac{TruePositives}{TruePositives + FalseNegatives}, \quad (4)$$

$$TNR = \frac{TrueNegatives}{TrueNegatives + FalsePositives}. \quad (5)$$

III. Implementation

A. Data Description

In this thesis, two datasets were considered, both provided by IPMA and containing time series. One dataset pertains to *E. coli* concentrations, consisting of multiple *Excel* files, each corresponding to a different shellfish production area. Each production area comprises one or more shellfish species and sampling locations. Although the date of the first and last *E. coli* measurements, as well as the number of measurements, are not uniform across all areas, in general, the *E. coli* concentrations spanned from January 2014 to December 2020. The other dataset consists of several *Excel* files containing daily measurements of meteorological variables from multiple meteorological stations along the Portuguese coast between

January 2015 and December 2020. The variables used in this thesis are described in Table II.

TABLE II: Description of all the variables from both datasets.

Variable	Description	Unit	Type
mean_temp	Mean air temperature	$^{\circ}C$	Numerical
max_temp	Maximum air temperature	$^{\circ}C$	Numerical
min_temp	Minimum air temperature	$^{\circ}C$	Numerical
mean_wind_intensity	Mean wind intensity	ms^{-1}	Numerical
mean_wind_dir	Mean wind direction	$^{\circ}$	Categorical
wind_dir	Wind direction (N, NE, ...)	—	Categorical
rainfall	Rainfall	mm	Numerical
E. coli	*E. coli* concentrations	MPN/100 g	Numerical

B. Data Preparation

Upon acquiring the data, a cleaning step consisting of correcting spelling errors was performed. Then, a *DataFrame*, a data structure that organises data into a 2-dimensional table of rows and columns, was created out of every species in each production area, meaning that each *E. coli* file originated as many *DataFrames* as species in that file. Therefore, each *DataFrame* contained the *E. coli* series corresponding to a single species in a single area. Moreover, to ensure an even distribution of the data, only one measurement per month was considered, which means that in months with multiple points, only the highest value of the concentration of *E. coli* was retained. Additionally, months without recorded measurements were marked as missing values in the *E. coli* series. Afterwards, *DataFrames* with few points (less than ten recorded *E. coli* concentrations) were discarded, and only some parts of the remaining ones were considered appropriate and were extracted: sets with ten or more samples with less than three missing values between each two were maintained. Figure 1 exemplifies this filtering step.

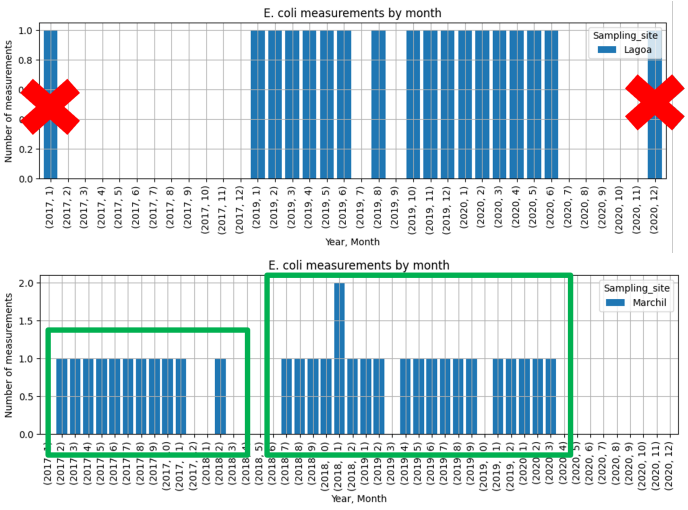

Fig. 1: Measurements of *E. coli* over time. (Top) The *E. coli* measurements with a red cross over them were removed; (Bottom right) There are two eligible sets of points. Since they are too far apart, the original *DataFrame* was replaced by two new ones, each consisting of one of the sets of points.

Then, to join all the data variables, each *DataFrame* was merged with the appropriate meteorological file, meaning that the files corresponding to the meteorological stations closest to each sampling location in each *DataFrame* were selected. The variables were then merged according to the dates of the *E. coli* series. In cases where the bacteria's concentration was recorded but some meteorological variables were missing, the average or mode of the three days preceding and following the date corresponding to the point in question was used depending on whether the variable at issue was numerical or categorical, respectively. If no value was present within this range, the corresponding point in the time series was marked as missing. In months without recorded *E. coli* measurements and, therefore, without values assigned to the meteorological variables, the mean/mode of the entire month from the most prevalent sampling site in the area was assigned to each meteorological variable on that month.

After integrating both datasets, the percentage of missing data was measured to assess which variables could be used further for forecasting. For each *DataFrame* if a particular variable contained more than 20% of missing values, that variable would be discarded when building the forecasting models. Otherwise, the missing values were filled in by a linear interpolation when building the models. Given the amount of data, the number of meteorological variables not discarded, and the number of points in each class of each *DataFrame*, three production areas were selected, meaning that the developed models were created to predict the concentration of *E. coli* in those areas. The chosen areas corresponded to the L1(3) and the L5b with the mussel species, as well as the RIAV1 with the cockle species.

C. Model Development

Since the ARIMA is a univariate model, it was only applied to the concentration of *E. coli*, meaning that the formed models attempted to predict future values of *E. coli* concentration using solely its past values. The data was split into training and test sets, comprising the first 80% and the remaining 20%, respectively. After splitting the data, linear interpolation was applied independently to both sets. Since the range of values of the *E. coli* series is immense, with maximum values around 9000 MPN/100g and minimum values around 18 MPN/100g, a threshold (2500 MPN/100g) was imposed to ceiling big concentrations. Due to the low amount of data, in between each two points of the training set, a gap was created and filled with a linear interpolation. This step was performed to increase the amount of data, and although it did not prove to be very relevant, it also did not negatively influence the forecasts, so it was maintained. Afterwards, four different models were created: ARIMA1 (set up to receive seasonal data and using the original values of the training data), ARIMA2 (non-seasonal and using the original values), ARIMA3 (seasonal and a logarithmic transformation was applied to the training data) and ARIMA4 (non-seasonal and a logarithmic transformation applied). The optimal parameters of each model (*p, d, q*) were determined with the function *auto_arima* from the package *pmdarima.arima*. This function was applied to the training set, and it conducted the Augmented Dickey-Fuller (ADF) test to determine the order of differencing *d*. Then,

different models were fitted for combinations of p and q, both ranging from 0 to 5. The *auto_arima* function determined the best model by finding the parameters combination that gave the lowest Akaike Information Criterion (AIC) value.

Since the VAR models are multivariate, and the meteorological variables spanned from 2015 onward, the data concerning the *E. coli* series preceding that year were discarded when building the models. Then, the Granger's Causality Test was conducted through the *grangercausalitytests* function from the *statsmodels.tsa.stattools* package to determine which variables provided predictive value for the *E. coli* prediction. No features showed predictive value for the *E. coli* variable in the L1(3) production area; in the RIAV1 area, the *mean_temp* and the *min_temp* were found to be relevant; and lastly, the *wind_dir* was the only significant variable in the L5b production area. Two VAR models were created: VAR1, trained with all variables, and VAR2, trained only with those considered essential by the Granger's Causality Test. Then, the data was split into training and test sets, and as for the ARIMA models, the missing values were handled with linear interpolation, every concentration of *E. coli* above 2500 MPN/100g was replaced by the value of 2500 MPN/100g and the amount of training data was increased. Since the VAR model requires the time series to be stationary, the ADF test was conducted on each series of the training set to verify that, and differencing was applied if necessary. Afterwards, the models were created, and the function *select_order* was used to select the appropriate order p of each model. This choice was made based on AIC, Bayesian Information Criterion (BIC) and Hannan-Quinn Information Criterion (HQIC).

For each area, one univariate and two multivariate Random Forest models were created in this thesis. Regarding the creation of the multivariate models, the same procedure as the one applied when creating the VAR models when it comes to the chosen variables was followed. Thereafter, a sliding window with length 4 was applied to the data in order to structure the time series into a supervised learning problem. After this transformation, the originated vectors were put together, and the resulting dataset was split into training, validation and test sets. Then, the same process of handling the missing values, increasing the number of samples and limiting the values was applied. After preparing the data, the models were built using the *RandomForestRegressor* from the *sklearn.ensemble* package and the models' hyperparameters were tuned, meaning that each model's performance in the validation set was compared for different combinations of hyperparameter values to determine the optimal combination. The options of values provided to the models for each of the tuned hyperparameters are presented in Table III.

Since ANNs require lots of data to learn a desired function and each area only contained a few points, every *DataFrame's* data corresponding to the same species as the one in the area under study was considered. Both univariate and multivariate ANN models were created, however, the latter did not employ the same variables as for the VAR case. Since the construction of the ANN models involved using data from several *DataFrames*, only the most common variables across them were considered. Therefore, all multivariate ANN models used the following variables: *mean_temp*, *max_temp*, *min_temp* and *rainfall*. Then, the values of each *DataFrame* were limited, and normalisation was applied to guarantee that variables with different scales could contribute equally to the model fitting. Afterwards, a sliding window was applied to each *DataFrame's* data; the resulting vectors were put together and split into training, validation and test sets. The test set remained the same as the other models, but the training and validation sets contained data from other *DataFrames* besides data from the area in question. Moreover, the amount of training data was increased. Each LSTM and FFNN layer was defined using the functions *LSTM* and *Dense*, respectively, and the optimal combination of hyperparameters was determined for each model. Table IV shows their options of values. This process was repeated for three different sliding window lengths (4, 12 and 19), accounting for three univariate and three multivariate LSTM and FFNN models for each selected area.

TABLE III: Value options of the tuned hyperparamenters for the Random Forest models.

Hyperparameter	Value Options
n_estimators	1000
max_features	*n_features*, $\sqrt{n\_features}$
max_depth	10, 32, 55, 77, 100
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 4
bootstrap	Yes, No

TABLE IV: Value options of the tuned hyperparamenters for the FFNN and LSTM models.

Hyperparameter	Value Options
Number of hidden layers	1, 2, 3
Number of neurons in each layer (FFNN)	128, 256, 512, 1024, 2048
Number of neurons in each layer (LSTM)	4, 16, 64, 128, 256
Dropout	0, 0.2, 0.4, 0.6
Activation function	*linear*, *tanh*, *ReLu*
Learning rate	1×10^{-2}, 1×10^{-3}, 1×10^{-4}
L1 regularization	0, 1×10^{-2}, 5×10^{-2}
L2 regularization	0, 1×10^{-2}, 5×10^{-2}

IV. Results

In this section, the results are presented and discussed. As mentioned before, the addressed problem can also be formulated as a classification one simply by labelling each prediction as part of the A or B class, depending on whether its value is below or above 230 MPN/100g.

A. ARIMA Models' Results

In each area, four ARIMA models were created, and their performance is presented in Table V. Looking at the RMSE values obtained in the training set, we can see that the ARIMA2 model outperformed the remaining ones in all areas. Overall, the results from the test sets are consistent with the ones from the training set, i.e., the ARIMA2 model was generally the best. In general, the results obtained by the ARIMA4 model, which, like the ARIMA2, was non-seasonal, were better than the ones by the ARIMA3, suggesting that the data was not seasonal.

TABLE V: Results obtained by the ARIMA models in each production area. The best values among the four models in each area for the training and test set are represented in bold.

Area	Model	Training Set	Test Set		
		RMSE	RMSE	TPR	TNR
L1(3)	ARIMA1	310.81	537.41	**1**	0.62
	ARIMA2	**303.66**	**535.22**	0.9	**0.88**
	ARIMA3	452.41	960.34	0.8	**0.88**
	ARIMA4	381.98	821.17	0.9	**0.88**
RIAV1	ARIMA1	612.10	703.99	**0.91**	0.4
	ARIMA2	**412.70**	**685.39**	**0.91**	0.4
	ARIMA3	676.53	1101.44	0.64	**0.6**
	ARIMA4	517.57	761.52	0.73	0.4
L5b	ARIMA1	216.56	551.02	**0.5**	**0.87**
	ARIMA2	**200.06**	491.49	**0.5**	**0.87**
	ARIMA3	264.99	**306.86**	**0.5**	**0.87**
	ARIMA4	251.16	433.72	0	**0.87**

B. VAR Models' Results

Two VAR models were created in each area, and their performance is shown in Table VI. Even though the VAR1 model consistently obtained lower RMSE values than the VAR2 model in the training set, the difference was not substantial, suggesting that the additional variables used by the VAR1 model did not provide significant advantages. In the RIAV1 area, the test set results are consistent with the training set, with the VAR1 model outperforming the other in all metrics. However, in the L5b production area, the opposite happened.

TABLE VI: Results obtained by the VAR models in each production area. The best values among the two models in each area for the training and test set are represented in bold.

Area	Model	Training Set	Test Set		
		RMSE	RMSE	TPR	TNR
L1(3)	VAR1	**282.57**	**595.98**	**0.7**	**0.75**
	VAR2	—	—	—	—
RIAV1	VAR1	**305.71**	**699.21**	**0.91**	**0.6**
	VAR2	315.12	709.05	0.73	**0.6**
L5b	VAR1	**170.02**	661.14	**0.5**	0.6
	VAR2	214.64	**568.48**	**0.5**	**0.8**

C. Random Forest Models' Results

Concerning the RF models, one univariate and two multivariate models were built to predict the concentration of *E. coli* in each area. The results obtained by the models in the validation set in each area are presented in Table VII. It should be noted that the univariate RF models contained more samples in the training data than the multivariate ones since the meteorological variables only spanned from 2015 onwards. This may have negatively contributed to the multivariate models' predictions. Overall, from Table VII, the univariate models outperformed the other ones in the validation set according to most metrics.

Finally, Table VIII provides the results obtained by the models in the test set, which, in general, are consistent with the ones from the validation set.

TABLE VII: Results obtained by the RF models in each production area. The best values among the three models in each area for the validation set are represented in bold.

Area	Model	RMSE	TPR	TNR	Transition Points	
					Correct Points	Count
L1(3)	URF	**551.80**	**0.75**	**0.79**	**3**	
	MRF1	772.76	**0.75**	0.5	2	4
	MRF2	—	—	—	—	—
RIAV1	URF	**891.68**	**0.67**	0.2	2	
	MRF1	940.59	0.5	0.4	**3**	8
	MRF2	967.99	0.5	**0.5**	2	
L5b	URF	**441.66**	**0**	0.75	**0**	
	MRF1	650.40	**0**	0.69	**0**	1
	MRF2	643.01	**0**	**0.81**	**0**	

TABLE VIII: Results obtained by the RF models in each production area. The best values among the three models in each area for the test set are represented in bold.

Area	Model	RMSE	TPR	TNR	Transition Points	
					Correct Points	Count
L1(3)	URF	**789.97**	0.6	**1**	**2**	
	MRF1	841.35	**0.9**	0.75	1	3
	MRF2	—	—	—	—	—
RIAV1	URF	**722.62**	**1**	0.2	**3**	
	MRF1	862.51	0.82	0	1	5
	MRF2	853.26	0.82	0	1	
L5b	URF	344.84	**0**	**0.8**	**0**	
	MRF1	**325.81**	**0**	0.67	**0**	4
	MRF2	328.20	**0**	0.67	**0**	

D. FFNN Models' Results

Regarding the FFNN models, the performance of the univariate and multivariate models in the validation set is presented in Tables IX and X, respectively. From these tables, the models in each area that outperformed the others according to most metrics were considered the best, which means that in the univariate case, the model UFFNN1 was selected in the L1(3) area and the UFFNN3 models were selected in the remaining areas. In the multivariate scenario, the same principle was applied: the MFFNN2 model in both the L1(3) and L5b areas. In the RIAV1 area, the MFFNN3 was considered the best multivariate FFNN model for predicting the concentration of *E. coli* since it presented TPR and TNR values less discrepant from each other.

TABLE IX: Results obtained by the univariate FFNN models in each production area. The best values among the three models in each area for the validation set are represented in bold.

Area	Model	RMSE	TPR	TNR
L1(3)	UFFNN1	**545.70**	**0.63**	0.83
	UFFNN2	572.49	0.52	**0.85**
	UFFNN3	552.48	0.61	**0.85**
RIAV1	UFFNN1	824.95	**0.71**	0.65
	UFFNN2	773.59	0.65	**0.71**
	UFFNN3	**728.35**	0.56	**0.85**
L5b	UFFNN1	631.11	0.77	0.74
	UFFNN2	606.52	0.76	**0.83**
	UFFNN3	**604.87**	**0.89**	0.59

Once the best FFNN models were selected, their performance in the test set was measured and is shown in Table XI.

Analysing the Table XI, it is evident that the multivariate

TABLE X: Results obtained by the multivariate FFNN models in each production area. The best values among the three models in each area for the validation set are represented in bold.

Area	Model	RMSE	TPR	TNR
L1(3)	MFFNN1	623.21	0.43	**0.88**
	MFFNN2	**590.40**	**0.62**	0.77
	MFFNN3	621.29	0.35	**0.88**
RIAV1	MFFNN1	999.62	0.57	**0.93**
	MFFNN2	810.42	**0.82**	0.53
	MFFNN3	**785.07**	0.67	0.81
L5b	MFFNN1	714.75	0.67	**0.76**
	MFFNN2	**625.91**	**0.83**	0.71
	MFFNN3	660.26	0.59	0.75

TABLE XI: Results obtained by the best FFNN models in each production area. The best values among the two models in each area for the test set are represented in bold.

Area	Model	RMSE	TPR	TNR	Transition points	
					Correct points	Count
L1(3)	UFFNN1	**727.40**	0.8	0.62	0	3
	MFFNN2	801.32	**1**	**0.75**	**2**	
RIAV1	UFFNN3	1209.36	0.36	**0.8**	**3**	5
	MFFNN3	**918.07**	**0.82**	0.2	2	
L5b	UFFNN3	482.71	0	**1**	**2**	4
	MFFNN2	**351.13**	**0.5**	0.8	1	

FFNN model in the L1(3) production area outperformed the univariate one. Regarding the other two areas, both the univariate and multivariate models struggled to achieve good results, leading to similar performances.

E. LSTM Models' Results

Following the same process as for the FFNNs, Tables XII and XIII correspond to the performances of the univariate and multivariate LSTM models in the validation set, respectively.

TABLE XII: Results obtained by the univariate LSTM models in each production area. The best values among the three models in each area for the validation set are represented in bold.

Area	Model	RMSE	TPR	TNR
L1(3)	ULSTM1	572.82	**0.73**	0.74
	ULSTM2	534.81	0.63	0.7
	ULSTM3	**515.81**	0.68	**0.81**
RIAV1	ULSTM1	810.50	0.71	0.71
	ULSTM2	815.86	0.81	**0.76**
	ULSTM3	**750.33**	**0.81**	0.72
L5b	ULSTM1	627.78	**0.83**	0.68
	ULSTM2	**622.27**	0.74	0.74
	ULSTM3	642.53	0.7	**0.83**

Similarly to the case of the FFNN models, for most cases, the LSTM models that outperformed the others in most metrics were considered the best models in each area. Therefore, in the RIAV1 area, the ULSTM3 and MLSTM2 were chosen. In the L1(3) area, the ULSTM1 and MLSTM1 models were chosen since each model's TPR and TNR were not as discrepant as they were for the other two models in both cases. In the L5b area, all univariate models presented similar results, therefore, the choice of the best model was arbitrary in favour of the ULSTM1, and in the multivariate scenario, the MLSTM2 was

TABLE XIII: Results obtained by the multivariate FFNN models in each production area. The best values among the three models in each area for the validation set are represented in bold.

Area	Model	RMSE	TPR	TNR
L1(3)	MLSTM1	604.92	**0.72**	0.79
	MLSTM2	601.13	0.61	**0.86**
	MLSTM3	**598.38**	0.62	0.84
RIAV1	MLSTM1	764.02	**0.9**	0.59
	MLSTM2	**671.68**	0.81	**0.67**
	MLSTM3	676.88	0.77	0.62
L5b	MLSTM1	666.10	0.82	0.78
	MLSTM2	**606.18**	**0.86**	0.67
	MLSTM3	636.54	0.6	**0.88**

chosen since it outperformed the other models in two of the three metrics. Once the best LSTM models were selected, their performance in the test set was measured and is shown in Table XIV.

TABLE XIV: Results obtained by the best LSTM models in each production area. The best values among the two models in each area for the test set are represented in bold.

Area	Model	RMSE	TPR	TNR	Transition points	
					Correct points	Count
L1(3)	ULSTM1	806.33	0.7	0.63	**0**	3
	MLSTM1	**728.34**	**0.8**	**0.75**	**0**	
RIAV1	ULSTM3	876.08	0.64	**0.6**	0	5
	MLSTM2	**835.95**	**0.82**	0.4	**1**	
L5b	ULSTM1	344.29	**0**	0.73	**0**	4
	MLSTM2	**332.66**	**0**	**0.8**	**0**	

In every area, the multivariate models outperformed the univariate ones regarding the RMSE score. In the L1(3) production area, both models achieved decent TPR and TNR values, with the MLSTM1 model providing better results according to all metrics. In the RIAV1 area, the MLSTM2 not only achieved a lower RMSE value than the univariate model but also labelled more B class points correctly, resulting in a higher TPR value. However, it struggled to label A class points, leading to a lower TNR score. In the L5b area, no model could adequately identify the points that belonged to the B class.

F. Model Comparison

In order to understand which model was the best in predicting the concentration of *E. coli* in each area, the results obtained by the best models in each area are shown in Tables XV to XVII.

Table XV shows that the autoregressive models achieved a lower RMSE value than the other models. Concerning the TPR, all models obtained a reasonable value ranging from 0.6 to 1. Achieving a good TPR means that the models were successful in accurately classifying points that belong to the B class, which warns producers of the need to purify the bivalves one month in advance. During this stage, the shellfish cannot be marketed for direct human consumption, so identifying such situations helps protect the population. Regarding the TNR, the univariate Random Forest model outperformed all the others, and the ARIMA2, as well as the multivariate models, obtained

TABLE XV: Results obtained by the best models in the test set of the L1(3) production area. The best values among the models are represented in bold.

Model		RMSE	TPR	TNR	Transition points	
					Correct points	Count
Univariate	ARIMA1	537.41	**1**	0.62	**2**	3
	ARIMA2	**535.22**	0.9	0.88	1	
	URF	788.97	0.6	**1**	**2**	
	UFFNN1	727.40	0.8	0.62	0	
	ULSTM1	806.33	0.7	0.63	0	
Multivariate	VAR1	595.98	0.7	0.75	1	
	MRF1	841.35	0.9	0.75	1	
	MFFNN2	801.32	**1**	0.75	**2**	
	MLSTM1	728.34	0.8	0.75	0	

a very acceptable value. Overall, the multivariate FFNN model seems to have been the best in terms of classification, with the highest TPR value, a very decent TNR value, and the ability to detect two of the three transition points. Regardless, both the ARIMA and the Random Forest models' performance was also good.

TABLE XVI: Results obtained by the best models in the test set of the RIAV1 production area. The best values among the models are represented in bold.

Model		RMSE	TPR	TNR	Transition points	
					Correct points	Count
Univariate	ARIMA2	**685.39**	0.91	0.4	**3**	5
	URF	722.62	**1**	0.2	**3**	
	UFFNN3	1209.36	0.36	**0.8**	**3**	
	ULSTM3	876.08	0.64	0.6	0	
Multivariate	VAR1	699.21	0.91	0.6	**3**	
	MFFNN3	918.07	0.82	0.2	2	
	MLSTM2	835.95	0.82	0.4	1	

In the RIAV1 area, the autoregressive models once again achieved the lowest RMSE values. Concerning the TPR, the URF model outperformed the other models, correctly labelling all B class points. Nevertheless, both the ARIMA2 and the multivariate models' performance according to this metric was good, with the minimum value being 0.82. Regarding the TNR, most models struggled to obtain a good value. Regardless, the UFFNN3, the ULSTM3 and the VAR1 models achieved a decent value, ranging from 0.6 to 0.8. Lastly, several models correctly identified three of the five transition points between classes. Thus, the VAR1 model is the best model for predicting the concentration of *E. coli* in the RIAV1 production area since it achieved one of the highest results in every metric.

TABLE XVII: Results obtained by the best models in the test set of the L5b production area. The best values among the models are represented in bold.

Model		RMSE	TPR	TNR	Transition points	
					Correct points	Count
Univariate	ARIMA3	**306.86**	**0.5**	0.87	1	4
	URF	344.84	0	0.8	0	
	UFFNN3	482.71	0	**1**	**2**	
	ULSTM1	344.29	0	0.73	0	
Multivariate	VAR1	661.14	**0.5**	0.6	**2**	
	VAR2	568.48	**0.5**	0.8	1	
	MFFNN2	351.13	**0.5**	0.8	1	
	MLSTM2	332.66	0	0.8	0	

Finally, regarding the L5b area, Table XVII provides the results of the best models created. Out of the three areas, this was the most challenging one. The main reason may be that the test set presented only two points that belonged to the B class. The idea was to see if the models, in a scenario characterised by low *E. coli* concentration values, could still detect sudden changes in the bacteria's concentration. Concerning the RMSE values, with the exception of the VAR models, the models obtained similar results. Regarding the TPR, the models either correctly labelled one of the two B class points or none, having the multivariate models been able to outperform the univariate ones according to this metric. When it comes to the TNR, the UFFNN3 model achieved the highest score, being able to label all the A class points properly. However, since its TPR was 0, we can conclude that the model assigned the A class label to every point in the test set. The models that achieved a proper TNR value and a TPR above 0 were the ARIMA3, VAR2 and MFFNN2 models. This area provided the biggest challenge, yet some models, namely the VAR, ARIMA3 and MFFNN2 models, overall achieved decent results.

The forecasts and respective confusion matrix of the best model in each area are displayed in Figures 2 to 4.

V. Conclusion and Future Work

Unforeseen reclassifications or prohibitions of shellfish production areas due to an increased concentration of faecal bacteria can pose a risk to human health and result in economic losses. Although the current strategy effectively protects consumers, it only responds after shellfish have been harvested, and fails to prevent the monetary losses of local producers, a challenging problem that has not been previously addressed along the Portuguese coast. In light of this, we constructed several machine learning forecasting models, both univariate and multivariate, to provide predictions of faecal bacteria contamination and address this issue. Autoregressive models (ARIMA and VAR), Random Forest models and ANNs (FFNNs and LSTMs) were built and used to predict the concentration of *E. coli* one month in advance in three selected shellfish production areas. The conducted experiments revealed that the simple autoregressive models (ARIMA and VAR) yielded good forecasting results, consistently achieving the lowest RMSE value. Furthermore, their performance classification-wise was also very good, overall being some of the best models for predicting the concentration of *E. coli* in some of the selected areas. Specifically, one VAR model was considered the best in the RIAV1 area according to the employed evaluation metrics, two ARIMA models were some of the best in the L1(3) area, and in the most challenging area, L5b, both ARIMA and VAR models outperformed the other models achieving decent results. Overall, the multivariate ANNs outperformed the univariate ones, even if only slightly. Accordingly, a multivariate FFNN model was one of the best in the L5b area, and one multivariate FFNN model surpassed the remaining models in the L1(3) area, achieving a TPR of 1 and a TNR of 0.75.

In general, the multivariate ANNs yielded slightly better results compared to their univariate counterparts. While this is promising, it may be beneficial to consider incorporating additional variables for predicting the concentration of *E. coli*.

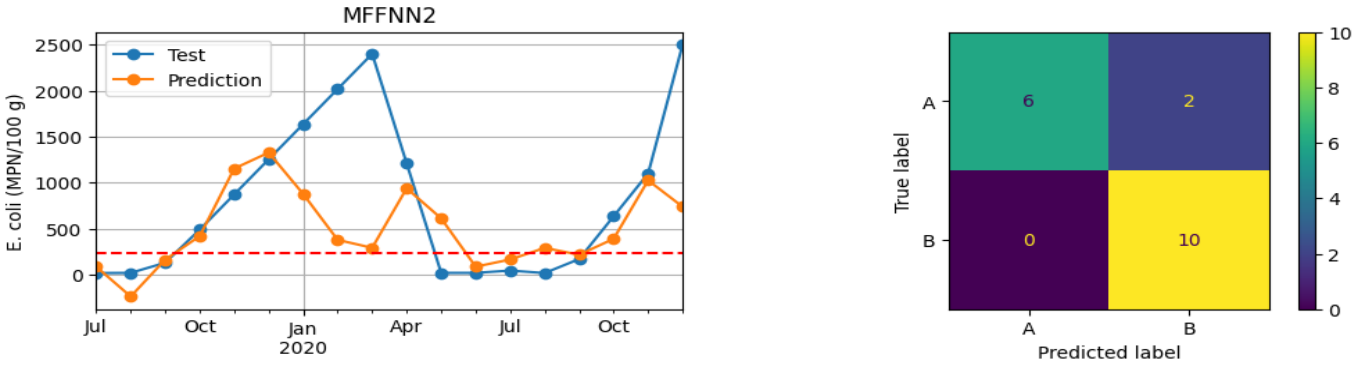

Fig. 2: MFFNN2 model's performance in the L1(3) area.

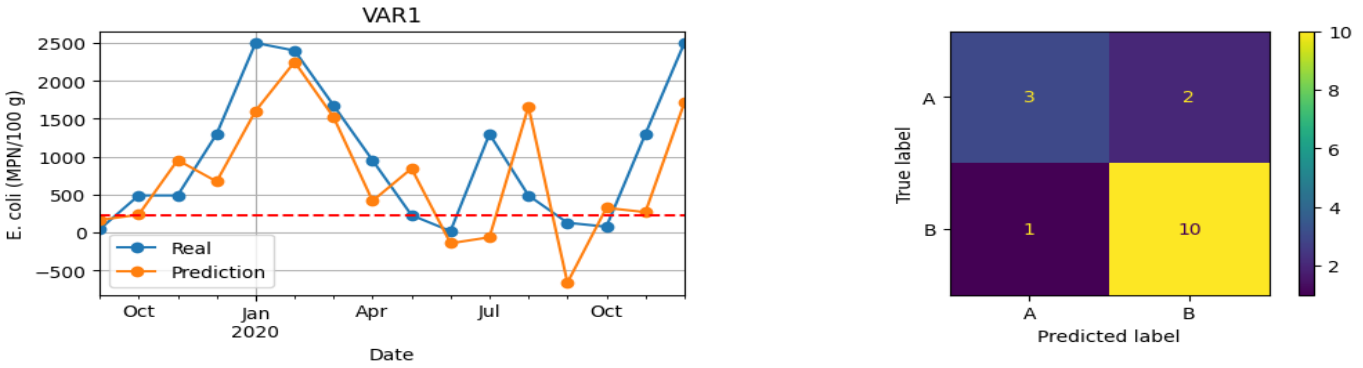

Fig. 3: VAR1 model's performance in the RIAV1 area.

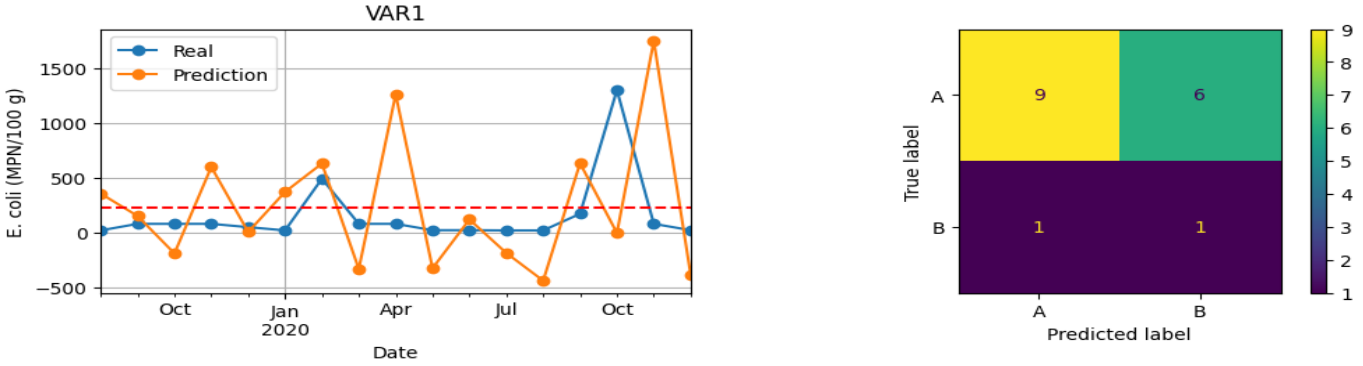

Fig. 4: VAR1 model's performance in the L5b area.

Examples are the ones mentioned in Section I-A which were identified as informative but were not available for this thesis. Although autoregressive models have produced reasonable results, it is plausible that the complexity of the addressed problem surpasses the capabilities of these simplistic models, and the available dataset might not be sufficient for achieving exceptional results with ANN models. Therefore, if more data were to become available, it would be worthwhile to revisit the study. Additionally, alternative models, such as convolutional neural networks (CNNs) and dynamic Bayesian networks (DBNs), could be explored to address this forecasting challenge.

Finally, it is worth mentioning that some results obtained while conducting this thesis led to the development of the following article: Ferraz, F., Ribeiro, D., Lopes, M. B., Pedro, S., Vinga, S., Carvalho, A. M. (2023). "Comparative analysis of machine learning models for time-series forecasting of *Escherichia coli* contamination in Portuguese shellfish production areas". In *The 9th International Conference on Machine Learning, Optimization and Data science – LOD 2023*. The article was submitted, accepted and presented at the LOD 2023 conference in Grasmere, Lake District, England, in September 2023.

Acknowledgment

I would like to thank my supervisors, Alexandra Carvalho and Susana Vinga, for all the support, guidance and feedback provided throughout this project. I would also like to thank

all co-authors of the article that this thesis originated for their guidance and feedback. I would like to acknowledge the Foundation for Science and Technology (FCT) funding through project MATISSE (DSAIPA/DS/0026/2019), and European Union's Horizon 2020 research and innovation programme under grant agreement No 951970 (OLISSIPO project). Finally, I would like to thank my family and friends, who always supported me throughout this entire journey.

REFERENCES

[1] Mateus, M.; Fernandes, J.; Revilla, M.; Ferrer, L.; Villarreal, M. R.; Miller, P.; Schmidt, W.; Maguire, J.; Silva, A.; Pinto, L.: "Early Warning Systems for Shellfish Safety: The Pivotal Role of Computational Science". In: *Computational Science - ICCS 2019*, pp. 361–375. Springer, (2019)

[2] Matarazzo Suplicy, F.: A review of the multiple benefits of mussel farming. Reviews in Aquaculture **12**(1), 204–223 (2020)

[3] Hallegraeff, G.; Anderson, D.; Cembella, A.; Enevoldsen, H.: Manual on Harmful Marine Microalgae. 2nd edition. UNESCO, (2004)

[4] Mok, J. S.; Shim, K. B.; Kwon, J. Y.; Kim, P. H.: Bacterial quality evaluation on the shellfish-producing area along the south coast of Korea and suitability for the consumption of shellfish products therein. Fisheries and Aquatic Sciences **21**(36), (2018)

[5] European Union: "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". In: *Official Journal of European Union*, 131 pp. 51–100., (2019)

[6] Schmidt, W.; Evers-King, H. L.; Campos, C. J. A.; Jones, D. B.; Miller, P. I.; Davidson, K.; Shutler, J. D.: "A generic approach for the development of short-term predictions of *Escherichia coli* and biotoxins in shellfish". In: *Aquaculture Environment Interactions*, vol. 10, pp. 173–185. (2018)

[7] Chen, Q.; Guan, T.; Yun, L.; Li, R.; Recknagel, F.: "Online forecasting chlorophyll a concentrations by an auto-regressive integrated moving average model: Feasibilities and potentials". In: *Harmful Algae*, Elsevier B. V., vol. 43, pp. 58–65, (2015)

[8] Bourel, M.; Segura, A. M.; Crisci, C.; López, G.; Sampognaro, L.; Vidal, V.; Kurk, C.; Pisccini, C.; Perera, G.: "Machine learning methods for imbalanced data set for prediction of faecal contamination in beach waters". In: *Water Research*, Elsevier B. V., vol. 202, (2021)

[9] Thoe, W.; Wong, S. H. C.; Choi, K. W.; Lee, J. H. W:"Daily prediction of marine beach water quality in Hong Kong". In: *Journal of Hydro-Environment Research*, Elsevier B. V., vol. 6, pp. 164–180, (2012)

[10] Cho, H.; Choi, U-J.; Park, H.: "Deep Learning Application to Time-Series Prediction of Daily Chlorophyll-a Concentration". In: *WIT Transactions on Ecology and the Environment*, vol. 215, pp. 157–163, (2018)

[11] Lee, S.; Lee, D.: Improved Prediction of Harmful Algal Blooms in Four Major South Korea's Rivers Using Deep Learning Models. International Journal of Environmental Research and Public Health **15**, (2018)

[12] Cruz, R. C.; Costa, P. R.; Krippahl, L.; Lopes, M. B.: Forecasting biotoxin contamination in mussels across production areas of the Portuguese coast with Artificial Neural Networks. Knowledge-Based Systems **257**, (2022)

[13] Anacleto, P.; Pedro, S.; Nunes, M. L.; Rosa, R.; Marques, A.: Microbiological composition of native and exotic clams from Tagus estuary: effect of season and environmental parameters. Marine pollution bulletin, **74**(1), 116–124, (2013)

[14] Jang, J.; Hur, H. G.; Sadowsky, M. J.; Byappanahalli, M. N.; Yan, T.; Ishii, S.: Environmental Escherichia coli: ecology and public health implications-a review. Journal of Applied Microbiology, **123**(3), 570–581, (2017)

[15] Campos, C. J. A.; Kershaw, S. R.; Lee, R. J.: Environmental Influences on Faecal Indicator Organisms in Coastal Waters and Their Accumulation in Bivalve Shellfish. Estuaries and Coasts. 36., 834-853, (2013)

[16] Brockwell, P. J.; Davis, R. A.: Introduction to Time Series and Forecasting. 2nd edition. Springer, (2002)

[17] Chatfield, C.: Time-Series Forecasting. CHAPMAN & HALL/CRC, (2001)

[18] Wei, W. W. S.: Multivariate Time Series Analysis and Applications. 1st edition. John Wiley & Sons Ltd, (2019)

[19] Cowpertwait, P. S. P.; Metcalfe, A. V.: Introductory Time Series with R. Springer, (2009)

[20] Tsay, R. S.: Multivariate Time Series Analysis: With R and Financial Applications. 1st edition. John Willey & Sons, (2014)

[21] Schonlau, M.; Zou, R. Y.: "The random forest algorithm for statistical learning". In: *The Stata Journal*, 20(1), 3–29, (2020)

[22] Haykin, S.: Neural Networks and Learning Machines. Pearson Prentice Hall, 3rd edition, (2008)

[23] Hochreiter, S.; Schmidhuber, J.: Long Short-term Memory. Neural computation, **9**, 1735-80, (1997)

[24] Goodfellow, I.; Bengio, Y.; Courville, A.: Deep Learning. MIT Press, (2016)

[25] Hewamalage, H.; Bergmeir, C.; Bandara, K.: Recurrent Neural Networks for Time Series Forecasting: Current status and future directions. International Journal of Forecasting **37**(1), 388–427 (2021)