Automation of Water Quality detection using Machine Learning

Fan Leon Wang [#1], Pyush Mittal Jain[*2]

[#]*Research Scholar, University of Munich*

Abstract - Coastal water quality management is a public health concern, as poor coastal water quality can potentially harbor pathogens that are dangerous to human health. Tourism-oriented countries need to actively monitor the condition of coastal water at tourist popular sites during the summer season. In this study, routine monitoring data of Escherichia Coli and enterococci across 15 public beaches in the city of Rijeka, Croatia, were used to build machine learning models for predicting their levels based on environmental parameters as well as to investigate their dynamics and relationships withenvironmental stressors. Gradient Boosting algorithms (Catboost, Xgboost), Random Forests, Support Vector Regression and Artificial Neural Networks were trained with routine monitoring measurements from all sampling sites and used to predict E: Coli and enterococci values based on environmental features. The evaluation of stability and generalizability with 10-fold cross validation analysis of the machine learning models, showed that the Catboost algorithm performed best with R2 values of 0.71 and 0.68 for predicting E: Coli and enterococci, respectively, compared to other evaluated ML algorithms including Xgboost, Random Forests, Support Vector Regression and Artificial Neural Networks.

Keywords: **coastal water quality, machine learning, shap, catboost, fecal indicator bacteria**

I. INTRODUCTION

Predicting the coastal water quality at public beaches would provide great benefits for the general human population as it is a major health issue due to a potential contamination with pathogenic microorganisms. Currently, in countries with a strong tourism sector, outdated information on coastal water quality is given because on-site measurements and laboratory analyses are time-consuming and lag behind the time frame in which a warning to the public would be necessary. For example, in Croatia, a country dependent on tourism in summer, sampling and laboratory analysis takes on average 2.2 days Furthermore, a lack of consistent criteria for estimating the potentially hazardous level of bacterial pollution is evident in the regulations between the EU and the US, with the EU directives using Escherichia Coli (EC) and enterococci (ENT) which are considered Faecal Indicator Bacteria (FIB) (Lušic et al., 2017; Directive, 2006), while the US regulations only consider ENT as the main indicator (USEPA, 2012). In previous studies, the Random Forest (RF) algorithm was used to predict FIB concentrations at 5 different beaches and showed that a logarithmic transformation of raw measurements increases the prediction accuracy of the ML model (Parkhurst et al., 2005). Also, Tree Regression (TR) and RF models were used to predict FIB values in freshwater, and it was found that precipitation is an important parameter for FIB prediction (Jones et al., 2013). Five different algorithms for FIB prediction on a single location using data collected 5 times a week for six bathing seasons were compared (Thoe et al., 2014). The trained models were auto-regressive and it was observed that the Artificial Neural Network (ANN) model performs good in terms of predicting the days when FIB exceeded levels of sufficient water quality. Gradient Boosting (GB) produces good results for FIB prediction at 7 different lake beaches (Brooks et al., 2016). The used data was collected for 3 bathing seasons, 2 to 4 times per week for 12-14 weeks per season. A hybrid wavelet auto-regressive ANN (WA-NAR) was used to predict EC concentrations at four different lake beaches and achieved high accuracy (Zhang et al., 2018). The study increased the EC measured data temporal resolution with the Monte Carlo Markov Chain (MCMC) method in order to increase the training and testing sets for the ML model. The collected data included measurements of EC and ENT which are prescribed by the European Union Bathing Water Directive, 2006/7/EC (EU BWD) (Directive, 2006). Data which was used in this study is a part of the Croatian national routine monitoring of coastal water quality which includes a time period from 2009 to 2020 for all locations except for KPW (measurements started in 2010) and KS (measurements started in 2019). A total of 10 measurements were made at every location for each bathing season using a set of thoroughly described methods (Lušic et al., 2017). Measurements were made every 15 days within the bathing season which started in mid-May and lasted until the end of September of each year and each measurement was made in the time span from 7 AM to 3 PM with most being between 8 AM and 9 AM. The Croatian criteria for coastal water quality assessment differs from the EU criteria for the measured EC level, while for ENT it is the same (Directive, 2008). In order for a beach to be considered safe for bathing activities, Croatian national rules prescribe that EC and ENT values must be below 300 CFU/100 mL and 185 CFU/100 mL, respectively, while the EU criteria defines values for EC of 500 CFU/100 mL for EC and for ENT of 185 CFU/100 mL, based upon a 90-percentile evaluation (Lušic et al., 2017). The Croatian national criteria applies to as an assessment for every sample and as a yearly/final evaluation, while the EU applies only as yearly/final evaluation of coastal water quality. In Fig. 1a, the percentage of EC measurements when a sampling location was classified as sufficiently safe and excellent can be observed, while Fig. 1b provides the same data for ENT measurements. A total of 120 measurements were made at every sampling location except for KPW where 110 measurements were made, and KS where the total number was 20. It is noticeable that the eastern part of the studied cluster, which includes locations KE, KW and KS, is the most problematic in terms of coastal water quality. However, with a total of 1690 measurements, it can also be concluded that the event of exceeding the criteria of being sufficiently safe for both EC and ENT is extremely rare at every sampling location.

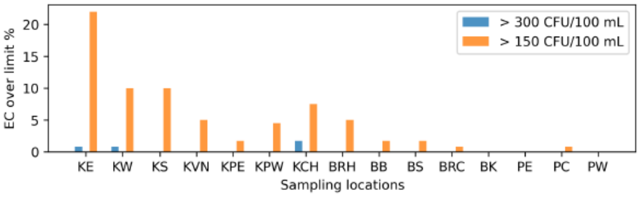


Fig 1a: Percentage of measurements of different samples

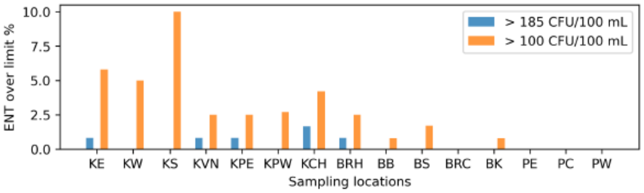
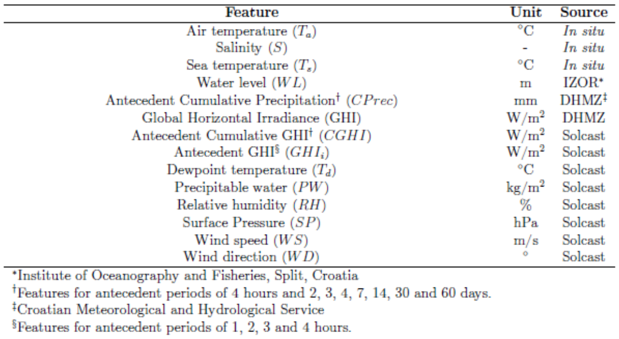


Fig 1b: Percentage of measurements of different samples

II. Environmental Parameters

Environmental parameters used as ML model input features included physical-chemical properties of the environment where and when the sample was taken and meteorological properties both at the time of sample acquisition and their antecedent cumulative values. A total of 33 features were considered for predictive modelling and are listed in Fig 3.



Feature	Unit	Source
Air temperature (T_a)	°C	*In situ*
Salinity (S)	-	*In situ*
Sea temperature (T_s)	°C	*In situ*
Water level (WL)	m	IZOR[*]
Antecedent Cumulative Precipitation[†] ($CPrec$)	mm	DHMZ[‡]
Global Horizontal Irradiance (GHI)	W/m^2	DHMZ
Antecedent Cumulative GHI[†] ($CGHI$)	W/m^2	Solcast
Antecedent GHI[§] (GHI_i)	W/m^2	Solcast
Dewpoint temperature (T_d)	°C	Solcast
Precipitable water (PW)	kg/m^2	Solcast
Relative humidity (RH)	%	Solcast
Surface Pressure (SP)	hPa	Solcast
Wind speed (WS)	m/s	Solcast
Wind direction (WD)	°	Solcast

[*] Institute of Oceanography and Fisheries, Split, Croatia
[†] Features for antecedent periods of 4 hours and 2, 3, 4, 7, 14, 30 and 60 days.
[‡] Croatian Meteorological and Hydrological Service
[§] Features for antecedent periods of 1, 2, 3 and 4 hours.

Fig 3: Environmental Features

Salinity, sea and air temperature were all measured at the sampling locations at the same time the EC and ENT data was collected. Water level data was obtained from the Institute of Oceanography and Fisheries (IZOR), Split, Croatia (IZOR, 2021) and it is a prediction generated by the XTide software (Flater, 1996). Water levels were linearly interpolated between two predicted water level points to exactly fit the time of sampling at each location. The precipitation data was obtained from the Croatian Meteorological and Hydrological Service (DHMZ) and the values were reported for every hour of the day. The precipitation data is used to calculate the cumulative values for antecedent intervals of 4 hours, and 2, 3, 4, 7, 14, 30 and 60 days before the exact time the sampling was made at each location. A total of 8 different features of precipitation data were created with the antecedent intervals for the ML model. The antecedent precipitation features are described in subsection 2.3.1. All other features were obtained at hourly intervals from Solcast, which is the solar resource assessment and forecasting data enterprise (Solcast, 2021). Solcast data has previously been validated and recommended for research purposes in the work by (Bright, 2019). All Solcast features were also linearly interpolated to match the sampling times. It should be noted that the air dewpoint temperature and relative humidity are recorded at 2 meters above ground level, the surface pressure accounts for the atmospheric pressure at ground level, while both wind speed and direction are recorded at 10 meters above ground level. The antecedent cumulative global horizontal irradiance (GHI) is calculated the same way and for the same antecedent intervals as cumulative precipitation, creating 8 additional features. Lastly, the values of GHI recorded 1, 2, 3 and 4 hours before the EC and ENT samples were collected were also considered as features (antecedent GHI). The created cumulative GHI and antecedent GHI features are described in subsection 2.3.1.

III. Predictive Modelling

ML Regression : The outputs of the ML model are the measured FIB values which correspond to each instance or state of the input features. In Fig. 4 the flowchart of the ML model is presented. Most features represent values at measurement time t, while CPrec and CGHI correspond to the antecedent intervals from measurement time t to th, where th 2 {4 hours, 2, 3, 4, 7, 14, 30, 60 days}, and the values of GHIi correspond to times ta (where ta 2 {1, 2, 3, 4 hours}) before the measurement time t. The ML algorithms used for EC and ENT values prediction were Gradient Boosting (Catboost, Xgboost), Random Forests, Support Vector Regression and Artificial Neural Networks (Multilayer Perceptron). Catboost (CB) is a gradient boosting toolkit which includes algorithmic advances such as the implementation of ordered boosting, a permutationdriven alternative to the classic gradient boosting algorithm, and an algorithm categorical features processing (Dorogush et al., 2018). The algorithm improvements made in Catboost are created to solve the prediction shift problem caused by target leakage. Xgboost (XGB) is a highly scalable tree boosting framework which incorporates a sparsity-aware algorithm and weighted quantile sketch for approximate tree learning (Chen and Guestrin, 2016). Xgboost combines cache access patterns, high data compression and sharding in order to build a tree boosting system.

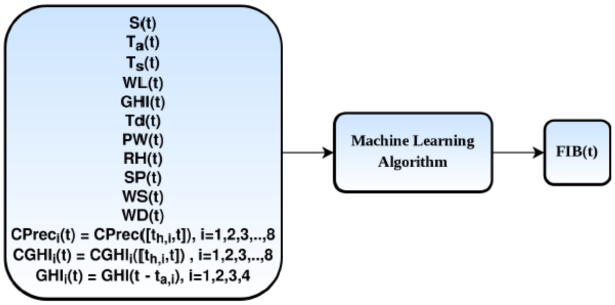

$S(t)$
$T_a(t)$
$T_S(t)$
$WL(t)$
$GHI(t)$
$Td(t)$
$PW(t)$
$RH(t)$
$SP(t)$
$WS(t)$
$WD(t)$
$CPrec_i(t) = CPrec([t_{h,i},t]), i=1,2,3,..,8$
$CGHI_i(t) = CGHI_i([t_{h,i},t])$, $i=1,2,3,..,8$
$GHI_i(t) = GHI(t - t_{a,i}), i=1,2,3,4$
Machine Learning Algorithm
FIB(t)

Fig 4: ML Regression flow chart

Random Forests (RF) is an ensemble ML algorithm which creatures multiple decision trees defined with features selected at random (Breiman, 2001). The decision trees with random features have both a lower variance and are less likely to cause overfitting of the ML model. A Multilayer Perceptron (MLP) Artificial Neural Network (ANN) consists of three basic layers (input, hidden and output) of artificial neuron nodes. Multiple hidden layers could be used in the MLP and each neuron in the hidden and output layers uses a nonlinear activation function which mimics the workings of human a brain (Nielsen, 2015). Support Vector Machines (SVM) are a class of supervised ML algorithms in which a superposition of kernel functions is used for data approximation (Drucker et al., 1997). Support Vector Regression (SVR) is a variation of SVM which is specifically used for to train ML regression models. Python 3.8. implementation of all algorithms was used in this study and Table 3 lists all of the algorithms used to train spatial and temporal FIB models. CB and XGB algorithms haven't been previously used to train FIB prediction models.

IV. Results & Discussions

In this section, an analysis of the accuracy of the ML model is assessed based on the coefficient of determination (R2), and the root mean squared error (RMSE), a standard evaluation metric for regression. The K-fold cross validation model sampling method with k=10 and data shuffling was used to investigate the robustness of the model and the model was trained and tested with all of the routine monitoring data at all of the sampling locations except for location KS (a total of 1670 instances). It can be seen that the most accurate algorithm in terms of both R2 and RMSE is CB and that it also showed robustness in its K-fold validation since the standard deviation values are the low for both EC and ENT models. The RF algorithm is a close second in terms of accuracy metrics, while the SVR prediction can be considered the least accurate with a moderate R2 score.

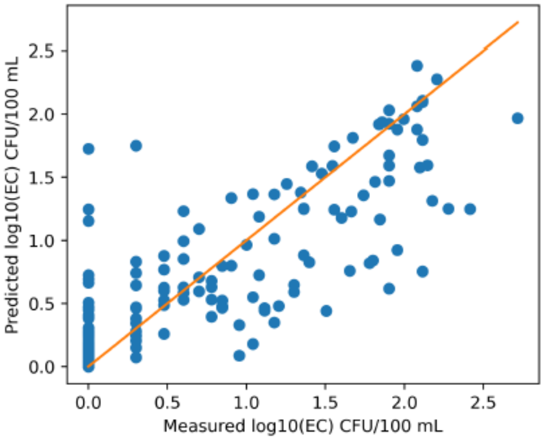


Fig 5: The correlation graph

The ENT model RMSE values are overall lower than those obtained by the EC model, however, the R2 metric is more useful in FIB predictive modeling than RMSE as it shows the level of correlation between the input features and the output variable.

ML Model Feature Interpretation: In this section, a feature analysis is presented using the game theoreticbased method SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017). Generally, the SHAP values generated by the method indicate a feature's control over a change in the model output and it is used for ML model interpretation. SHAP is a powerful tool that has been recently used in various cutting edge research areas such as explaining the ML prediction of hypoxaemia prevention during surgery (Lundberg et al., 2018), interpreting the ML modelbased behavior of nanophotonic structures (Yeung et al., 2020), and explaining the relationship between the stream water quality (which includes EC measurements) with urban development patterns (Wang et al., 2021b). The SHAP approach is used to interpret the output of any ML model, but in this specific study it is used to interpret how the input features are related to the output values of the most accurate - CB model with a 90/10 training/testing data split. The feature with the highest mean SHAP value contributes the most to the model output (which are the EC and ENT values in this study), and therefore has the highest predictive power.

Conclusion:

In this research, EC and ENT routine monitoring data collected at 15 different sampling locations in the city of Rijeka, Croatia, were used to create spatial and temporal ML models for the purpose of predicting FIB values based on environmental features, and getting a better insight into the dynamics and correlations between FIB and environmental variables. The main contributions of this study can be ummarized with the following points: both EC and ENT ML models were created with the routine monitoring data (colllected at all but one location) and environmental features. Gradient Boosting algorithms (Catboost, Xgboost), Random Forests, Artificial Neural Networks and Support Vector Regression were used to train and test the EC and ENT models. Cross validation showed that the Catboost algorithm performs the best with an achieved strong R2 score of 0.71 for EC prediction and 0.68 for ENT. The Random Forest algorithm achieved the second best score for all trained EC and ENT ML models. Secondly, a spatial ML EC and ENT prediction was done on one location to investigate the accuracy of the model in that regard. It was found that the model trained with the Catboost algorithm performs exceptionally well, achieving R2 scores of 0.85 and 0.83 for EC and ENT predictions, respectively. A spatial model could show its usefulness in finding the locations of groundwater springs as they are severely linked with EC and ENT through salinity.

References

[1] Alkan, U., Elliott, D., Evison, L., 1995. Survival of enteric bacteria in relation to simulated solar radiation and other environmental factors in marine waters. Water Research 29, 2071–2080.

[2] Avila, R., Horn, B., Moriarty, E., Hodson, R., Moltchanova, E., 2018. Evaluating statistical model performance in water quality prediction. Journal of Environmental Management 206, 910–919.

[3] Benac, C., Rubinic, J., Ožanic, N., 2003. The origine and evolution of coastal and submarine springs in bakar bay. Acta carsologica 32.

[4] Biondic, B., Dukaric, F., Kuhta, M., Biondic, R., 1997. Hydrogeological exploration of the rjecina river spring in the dinaric karst. Geologia Croatica 50, 279–288.

[5] Thara, D.K., Premasudha, B.G., Nayak, R.S. et al. Electroencephalogram for epileptic seizure detection using stacked bidirectional LSTM_GAP neural network. Evol. Intel. 14, 823–833 (2021). https://doi.org/10.1007/s12065-020-00459-9

[6] Bonacci, O., Oštric, M., Bonacci, T.R., 2018. Water resources analysis of the rjecina karst spring and river (dinaric karst). Acta Carsologica 47.

[7] Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

[8] Bright, J.M., 2019. Solcast: Validation of a satellite-derived solar irradiance dataset. Solar Energy 189, 435–449.

[9] Thara D.K., PremaSudha B.G, Fan Xiong, Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques, Pattern Recognition Letters, Volume 128, 2019, Pages 544-550, ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2019.10.029.

[10] Brion, G.M., Lingireddy, S., 1999. A neural network approach to identifying non-point sources of microbial contamination. Water research 33, 3099– 3106.

[11] Brooks, W., Corsi, S., Fienen, M., Carvin, R., 2016. Predicting recreational water quality advisories: A comparison of statistical methods. Environmental Modelling & Software 76, 81–94.

[12] Byappanahalli, M.N., Nevers, M.B., Korajkic, A., Staley, Z.R., Harwood, V.J., 2012. Enterococci in the environment. Microbiology and Molecular Biology Reviews 76, 685–706.

[13] Thara D.K., PremaSudha B.G., Fan Xiong, Epileptic seizure detection and prediction using stacked bidirectional long short term memory, Pattern Recognition Letters, Volume 128, 2019, Pages 529-535, ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2019.10.034.

[14] Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794.

[15] Davies-Colley, R.J., Bell, R.G., Donnison, A.M., 1994. Sunlight inactivation of enterococci and fecal coliforms in sewage effluent diluted in seawater. Applied and environmental microbiology 60, 2049–2058.

[16] Directive, E.B.W., 2006. Directive 2006/7/ec of the european parliament and of the council of 15 february 2006 concerning the management of bathing water quality and repealing directive 76/160. EEC .

[17] Dorogush, A.V., Ershov, V., Gulin, A., 2018. Catboost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363 .

[18] Drucker, H., Burges, C.J., Kaufman, L., Smola, A., Vapnik, V., et al., 1997. Support vector regression machines. Advances in neural information processing systems 9, 155–161.

[19] Družeta, S., Ivic, S., 2020. Indago - python module for numerical optimization. https://pypi.org/project/Indago/. Accessed: July 03, 2021.