Predictive modeling of microbiological seawater quality in karst region using cascade model

Lučin, Ivana; Družeta, Siniša; Mauša, Goran; Alvir, Marta; Grbčić, Luka; Lušić, Darija Vukić; Sikirica, Ante; Kranjčević, Lado

Source / Izvornik: **Science of The Total Environment, 2022, 851**

Journal article, Accepted version
Rad u časopisu, Završna verzija rukopisa prihvaćena za objavljivanje (postprint)

https://doi.org/10.1016/j.scitotenv.2022.158009

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:184:479027

Rights / Prava: Attribution-NonCommercial-NoDerivatives 4.0 International/Imenovanje-Nekomercijalno-Bez prerada 4.0 međunarodna

Download date / Datum preuzimanja: **2024-05-17**

Repository / Repozitorij:

Repository of the University of Rijeka, Faculty of Medicine - FMRI Repository

UNIRI DIGITALNA KNJIŽNICA

dabar
DIGITALNI AKADEMSKI ARHIVI I REPOZITORIJI

arXiv:2202.05664v1 [cs.LG] 11 Feb 2022

Predictive modeling of microbiological seawater quality classification in karst region using cascade model

Ivana Lučin[a,c], Siniša Družeta[a,c], Goran Mauša[b,c], Marta Alvir[a], Luka Grbčić[a,c], Darija Vukić Lušić[c,d,e], Ante Sikirica[a,c], Lado Kranjčević[a,c,*]

[a]*Department of Fluid Mechanics and Computational Engineering, Faculty of Engineering, University of Rijeka, Vukovarska 58, Rijeka, 51000, Croatia*
[b]*Department of Computer Engineering, Faculty of Engineering, University of Rijeka, Vukovarska 58, Rijeka, 51000, Croatia*
[c]*Center for Advanced Computing and Modelling, University of Rijeka, Radmile Matejčić 2, Rijeka, 51000, Croatia*
[d]*Department of Environmental Health, Faculty of Medicine, University of Rijeka, Braće Branchetta 20/1, Rijeka, 51000, Croatia*
[e]*Department of Environmental Health, Teaching Institute of Public Health of Primorje-Gorski Kotar County, Krešimirova 52a, Rijeka, 51000, Croatia*

Abstract

In this paper, an in-depth analysis of *Escherichia coli* seawater measurements during the bathing season in the city of Rijeka, Croatia was conducted. Submerged sources of groundwater were observed at several measurement locations which could be the cause for increased *E. coli* values. This specificity of karst terrain is usually not considered during the monitoring process, thus a novel measurement methodology is proposed. A cascade machine learning model is used to predict coastal water quality based on meteorological data, which improves the level of accuracy due to data imbalance resulting from rare occurrences of measurements with reduced water quality. Currently,

*Corresponding author
Email address: lado.kranjcevic@riteh.hr (Lado Kranjčević)

the cascade model is employed as a filter method, where measurements not classified as excellent quality need to be further analyzed. However, with improvements proposed in the paper, the cascade model could be ultimately used as a standalone method.
Keywords: bathing water quality, machine learning, fecal pollution, cascade prediction modelling, karst region

---

1. Introduction

Microbiological contamination presents a great concern in areas where water bodies are used for recreational activities since the existence of pathogens can cause serious health problems (Solo-Gabriele et al., 2016). This is especially important for tourism-oriented countries, such as Croatia, since bathing locations with high water quality can attract tourists, and maintenance of such favourable repute is one of the main priorities. Currently, the main methodology for water quality classification consists of fortnightly measurements of feacal indicator bacteria, such as *Escherichia Coli* or enterococci. Unfortunately, measurements are temporary and spatially sparse as they are expensive and time consumable. This is a considerable problem since studies observed that number of microbes has a high temporal and spatial variation (Ekklesia et al., 2015; Vukić Lušić et al., 2017). Additionally, currently in Croatia, sampling and laboratory testing take about 2.5 days, thus the information is already outdated by the time it is obtained. Therefore, methods for predicting water quality integrating meteorological data are increasingly

being investigated.

Many factors need to be taken into consideration when investigating microbe concentrations in the seawater, such as solar radiation, tides, wind intensity and direction, rainfall, the density of bathers, presence of rivers and canals near the bathing area, etc. (Agency, 2009; He and He, 2008; Cho et al., 2010). In a number of studies, rainfall was deemed greatly influential for microbiological contamination both in coastal (Dwight et al., 2011; Vukić Lušić et al., 2017; He et al., 2019) and underground waters (Knierim et al., 2015; Mance et al., 2018; Buckerfield et al., 2019). A more detailed overview of numerical modeling approaches and the influence of normal and extreme storm events on *E. coli* values in coastal waters was reviewed in Weiskerger and Phanikumar (2020).

When investigating the water quality of coastal areas, specifics of each location must be taken into consideration. He et al. (2019) conducted an analysis of two beaches, approximately 20 km apart in China, after a single storm event where a difference in water quality was observed due to distinct beach environments. Viau et al. (2011) investigated bacterial pathogens in Hawaiian coastal streams which were shown to be pollution sources associated with beach locations. In Kucuksezgin et al. (2019) the enclosed bay of Izmir Bay, Turkey was analyzed, where domestic and industrial wastes contribute to reduced water quality. In Verga et al. (2020) an analysis of seasonal and spatial variability of water quality in Patagonia, Argentina was conducted. Observed problem with sewage and draining systems was linked to insuffi-

cient dilution of wastewater in seawater. Chahouri et al. (2021) conducted an assessment of both beach and estuary location in Agadir bay, South-West Morocco. It is noted that estuary location is influenced by effluents from untreated wastewater and agricultural activities and the tourist beach is center of human activities, where greater fecal streptococci loads were observed at the estuary location. These studies indicate need for consideration of both urban development and geographical specifics of each location.

The terrain of the Croatian coast is mostly of karst type and characterized by high porosity and numerous subterranean channels, which makes it very vulnerable to pollution since surface water can quickly and easily enter groundwater (Pikelj and Juračić, 2013). Investigation of *E. coli* pollution in karst was conducted in a number of studies (Davis et al., 2005; Laroche et al., 2010). Sources of groundwater contamination can include landfills (Kogovšek and Petrič, 2013), sewage outflows (Heinz et al., 2009; Stange and Tiehm, 2020), agricultural or urban land-use type (Reed et al., 2011; Buckerfield et al., 2019), etc. For this reason, the water quality of karst aquifers used for drinking water is regularly monitored with special care, while karst aquifers not used for drinking water are not regularly monitored due to their reduced importance.

With growing interest in increased protection of water surfaces, prediction modeling is being used to provide information to the general public regarding potential health risks. He and He (2008) used Artificial Neural Network (ANN) to predict water quality regarding stormwater runoff with

reported less than 10% false positive or negative rates. In de Souza et al. (2018b) and de Souza et al. (2018a) regression models for predicting fecal indicators in coastal waters are developed and optimized, where cumulative solar radiation and cumulative rainfall values were shown to greatly influence the prediction of fecal pollution. He et al. (2019) used Multiple linear regression (MLR) model to predict pathogen contamination using environmental data collected during the storm event. Grbčić et al. (2021) investigated the efficiency of different machine learning algorithms to predict E. Coli and enterococci values based on environmental features.

The main premise of the proposed work is that a prediction model can be created which would use meteorological data for predicting *E. coli* values, by use of which forecasts could be made and warnings to the general public can be given in advance. The main contribution to this goal was made by utilising Random Forest classifier to predict water quality based on meteorological data. Data used for model training is obtained from available measurements of water quality during the bathing seasons 2009-2020 for Rijeka, Croatia. An investigation of different clustering methods of bathing locations was conducted with the addition of feature analysis. Additionally, a novel cascade prediction model framework, aimed at classifying measurements as excellent water quality, is proposed. Due to comprising a series of prediction models, it enables model fine-tuning for different physical processes with increased prediction accuracy. The proposed model provides great flexibility and as such can be used on pollution measurement datasets with sparse cases of

high pollution, which are the most common. In-depth analysis of measured data indicated the potential influence of submerged coastal springs which are specific for karst soil. In previous research (Vukić Lušić et al., 2017), these springs were not considered as possible pollution sources because they predominantly dry up during the bathing season. However, during the investigation of hydrogeological data for the monitored region, it was found that some of them are active throughout the whole year. Considering these new findings further research directions are presented in the discussion section.

2. Materials and methods

2.1. Data collection

The city of Rijeka is located in Kvarner Bay and is the third-largest Croatian city with important industrial locations such as shipyard and port, but with the increasing tendency to become a recognized touristic location. The city has Mediterranean climate, with dry and warm summers, and its surrounding is also characterized by a large amount of rainfall due to the proximity of Dinaric Alps. In the city of Rijeka precipitation is estimated at about 1540 mm per year, with 550 mm per year for period May to September. The average temperature of the warmest months (July and August) is 23.1°C (Zaninović et al., 2008). The bathing season usually lasts from mid-May until the end of September, and in that period regular measurements of water quality are conducted. Measurements from 9 locations on the west side of the city are considered for analysis with mentioned locations spread

over roughly 2 km length of the coastline. Locations of the measurement points can be observed in Figure 1. Regular measurements, with fortnightly intervals, from 2009 to 2020, are used as data inputs for the prediction model. Samples were taken from the boat in a short timeframe. Samples from all considered measuring points were collected, on average, within 30 minutes. Consequently, differences in atmospheric conditions are very small or nonexistent, between measurements taken on the same day. For each measurement point, samples are taken at a similar distance (20 meters) from the coast. During several months in 2012 and 2014, additional measurements were conducted every 4 hours for 5 measuring points: KBW, KBE, KW, KE, and 3M (see Figure 1). Additional measurements taken in the morning, which is the time when regular measurements are conducted, were also included in the analyzed dataset to increase the number of measurements. Additional measurements were always taken from the same location at the coast. Water temperature, salinity, and air temperature were measured in situ and *E. coli* value was analyzed in the laboratory from the collected sample. For *E. coli* enumeration membrane filtration technique was used, according to the ISO-9308-1 method for period 2009-2017 and temperature-modified ISO-9308-1 method for period 2018-2021. Cultivation was performed on CCA nutrient media (Chromogenic Coliform agar, Biolife Italiana S.r.l., Milan, Italy) for 4 h at 36 $\pm$ 2°C followed by 20 h incubation at 44 $\pm$ 0.5°C (Jozić and Vukić Lušić, 2018; Jozić et al., 2018). Further details on data collection and analysis are given in Vukić Lušić et al. (2017).

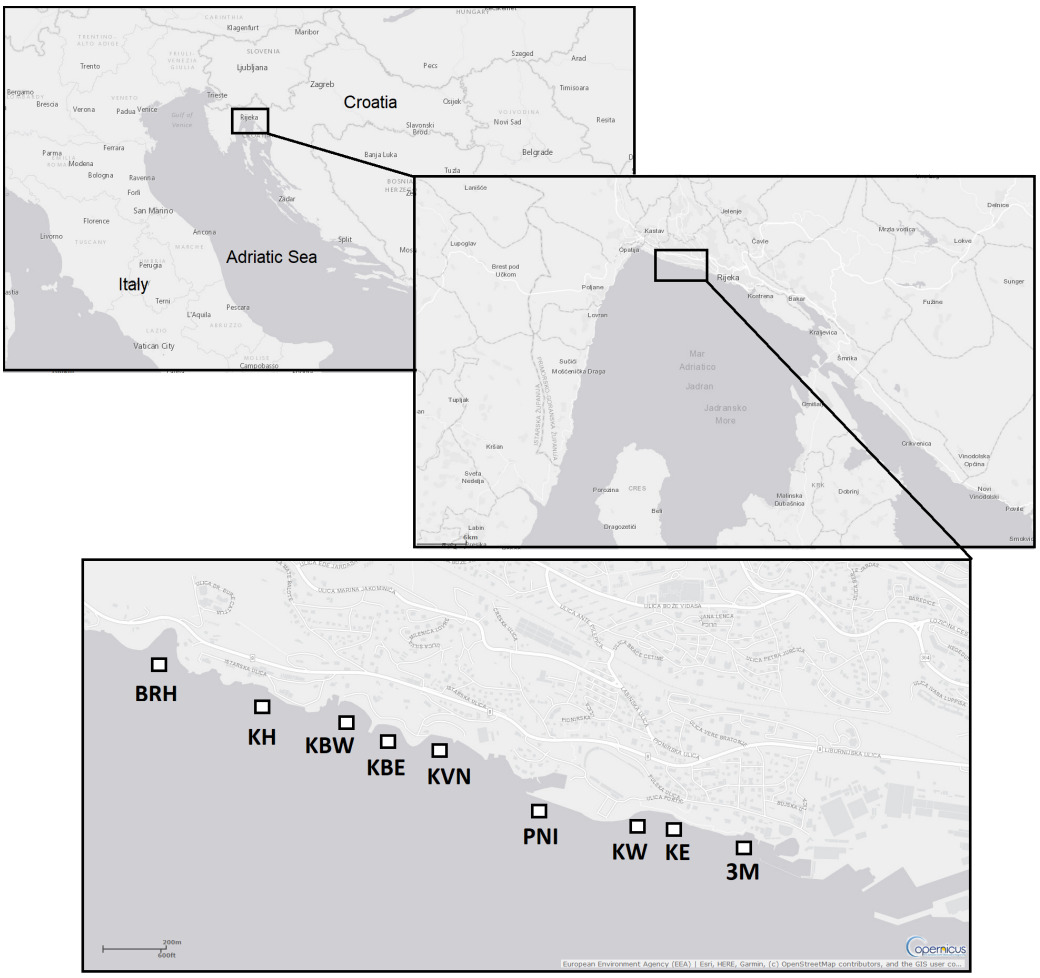

Figure 1: Measurement locations in Rijeka Bay (Kvarner Bay, Croatia) (European Union, 2018).

In total, considered measurements have 1137 records. It must be noted that not all measuring points have the same number of measurements, since 5 locations (KBW, KBE, KW, KE, and 3M) have additional measurements and measurements for PNI started in 2019.

2.2. Environmental parameters

Features chosen for prediction modelling were meteorological data obtained during measurements: water temperature, air temperature, and salinity. Additionally, solar irradiance as current Global Horizontal Irradiance (GHI) and cumulative irradiance for the past 4 hours are also considered. The chosen data was taken from Solcast (2021) database. The rainfall data was obtained from the Croatian Meteorological and Hydrological Service (DHMZ). Different combinations of cumulative rainfall values, such as previous 24 hours, previous 48 hours, etc., were also considered since in a number of previous studies influence of rainfall, especially storm events, were investigated (He et al., 2019; Weiskerger and Phanikumar, 2020). It was observed that a very small number of measurements have any rainfall from the previous several days, thus cumulative sums from $4-7$ and $7-14$ days are considered as a possible indication of soil saturation, which can happen if a larger amount of rain is present during a longer period of time. If soil is saturated, new rain can influence the activation of underground sources in the sea, which can increase the amount of *E. coli*.

2.3. Data analysis and preparation

The considered measurement points are chosen for prediction modeling since, historically, these locations have lower water quality than other bathing locations in the city, even though only a small amount of these measurements show less than excellent water quality. EU legislation (EC, 2006) prescribe

two fecal indicator bacteria, *E. coli* and enterococci where in Vukić Lušić et al. (2017) it was observed that *E. coli* criteria in Croatian legislation (Directive, 2008) are more stringent. Thus in this work, the prediction models are investigated considering only *E. coli* limits. Bathing water quality can be divided into three categories, where, by Croatian standards, excellent water quality is for *E. coli* measurements in the range 0-150 CFU/100 mL (in at least 95% of samples), good water quality in *E. coli* range 150-300 CFU/100 mL in at least 95% of samples, and sufficient quality for the range up to 300 CFU/100 mL in at least 90% of samples. EU criteria allow *E. coli* values up to 250 CFU/100 mL (in at least 95% of samples) for excellent water quality, good water quality in the range 250-500 CFU/100 mL (in at least 95% of samples) and sufficient water quality for *E. coli* values up to 500 CFU/100 mL in at least 90% of samples. If Croatian criteria is applied, and value for *E. coli* contamination is taken as 300 CFU/100 mL, there are 21 records in the dataset with *E. coli* values above that threshold. If contamination measure is taken as 150 CFU/100 mL, 122 records have *E. coli* value above that threshold. It can be observed that there is only 11% of less-than-excellent and only 1.8% of less-than-sufficient water quality measurements. Thus, an in-depth analysis of these cases is performed. Histogram of considered measurements can be observed in the Figure 2.

It was observed that greater values of *E. coli* can be roughly grouped in three types of pollution events. The first are unexpected and unexplained surges in *E. coli* concentration during the monitoring. They are followed

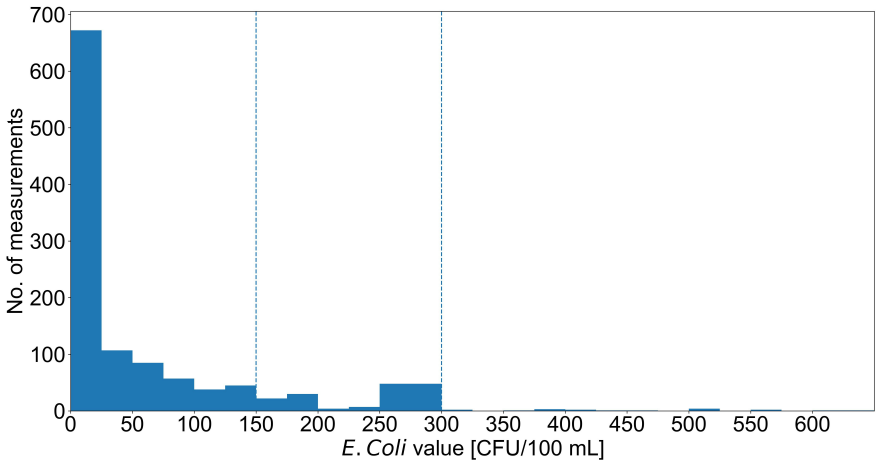

Figure 2: Histogram of *E. coli* measurements. Vertical lines indicate limits for excellent and sufficient water quality.

by control measurements and are considered as an accidental event. These measurements are considered outliers since they show no correlation with meteorological data, thus it is reasonable to believe that these are singular incident events. Therefore these measurements are removed from the data set to improve the prediction model performance. After this correction, 1133 measurement remained, with 17 records having *E. coli* value above 300 CFU/100 mL (1.5% of measurements) and 118 records above 150 CFU/100 mL (10.4% of measurements). The second group of events are typically occurring during spring, where it was observed that water salinity is reduced in all measuring points, which can be explained by the greater amount of precipitation during the spring period which could lead to greater *E. coli* concentrations. The average salinity value for the period from May to mid-

June was 30.6 with an average *E. coli* value 72.3 CFU/100 mL, and for the period from mid-June to September, average salinity was 34 with an average *E. coli* value 46 CFU/100 mL, which supports the presented assumption. The third type of events are the occurrences where it was observed that some measuring points show higher *E. coli* value coupled with lower water temperature and lower water salinity than other nearby measuring points, during the same measurement period. This indicates the presence of local sources of groundwater at these measuring points. It must be noted that not all measurements with lower water quality can be put in these categories, thus prediction modelling needs to be utilized to find additional correlations.

The mean value of salinity was analysed for all measuring points to investigate the possible correlation with *E. coli* value and results are presented in Table 1. It can be observed that locations with greater average *E. coli* value also have smaller average salinity. This can be explained by the fact that underground water sources, which presence can be observed through reduced water salinity, collect bacteria from the watershed area and transport them into the seawater. However, this assumption needs to be further investigated.

2.4. Random Forest classifier

Machine learning algorithms are designed to find an underlying correlation or patterns between the data input and data output to provide a prediction for unseen data. Machine learning algorithms can be divided into regression and classification algorithms, where the first group of algorithms

Table 1: Salinity and *E. coli* characteristics of measuring points.

Measuring point	*E. coli* (CFU/100 mL)		Salinity	
(number of measurements)	Mean	Median	Mean	Median
BRH (122)	36	5	35	36
KH (123)	47	7	34.5	35.7
KBW (144)	36	7	34.7	35.9
KBE (149)	26	5	34.8	36
KVN (119)	35.8	8	34.4	35.5
PNI (20)	56.8	26	31.8	34.5
KW (151)	78	35	31.6	33.3
KE (155)	86.4	60	30	32.2
3M (150)	72.3	25	30.5	32.9

aims to predict the exact value of the output variable, while the other try to separate data into predefined groups. Since the problem considered in this paper, by nature of corresponding regulation, deals with water quality groups, a classifier algorithm was considered for the prediction of water quality. Prediction models were constructed with only two classes, i.e. a prediction is made whether *E. coli* value is above or below the considered limit. Random Forest classifier implementation in the Python library Scikit-learn (Pedregosa et al., 2011) version 0.20.3 was used.

Random Forest classifier is an ensemble type of machine learning algorithm which was first proposed by Breiman (2001). It consists of multiple decision trees which stand as independent prediction models. The bootstrap method is used to provide a unique subset for the training of each decision tree while the aggregation method is used to count the class with the most prediction occurrences which is then considered as the prediction of the Random Forest model.

Different combinations of Random Forest Classifier parameters were investigated, where best results were obtained for 100 estimators, maximum depth of 10, and the minimum number of samples required to split an internal node equal to 6. All other parameters were kept at default values. Considered parameters were obtained for prediction model trained on the first group of uniformly distributed data with *E. coli* classification limit 150 CFU/100 mL which is Croatian criteria for excellent water quality.

2.5. Prediction models and sampling methodology

Based on different data clustering strategies, three different prediction models were created. One where all measurements were taken as inputs for a single prediction model since all measuring points are geographically near each other. The second and third models were constructed for westernmost 5 and easternmost 4 measuring points respectively, where measuring points with similar mean salinity values are grouped. The reasoning behind it is that sources of groundwater considerably influence *E. coli* value, where physical processes for locations with and for locations without those sources can be considered different. If that premise is true, a single prediction model cannot successfully predict for both considered behaviours.

The available data was split into two subsets: 80% data for training and 20% for testing. Since it is observed that less than 2% of measurements have contamination levels above the regulation limit, it is expected that random split of training and testing data, such as k-fold analysis, would greatly in-

fluence prediction model accuracy, e.g. it is possible that all contaminated measurements end up being sorted in the training set, resulting in high accuracy of prediction model trained on testing set with no measurements above the limit. To take that into account, six different dataset splits of training and testing data were investigated. For the first two datasets, all measurements are sorted by *E. coli* value, and each fifth measurement is taken for the testing set and remaining measurements are used for model training. In this way, the same ratio of measurements above the considered limit is maintained both for the training and testing set. To take into account temporal distribution, two different years are extracted from the dataset so as to serve as test sets, one with smaller and one with a greater number of measurements with reduced water quality where remaining measurements are used as the training set. Similarly, to take into account spatial distribution, two different measuring points are considered for prediction, one with a smaller average *E. coli* value and one with a greater average *E. coli* value. For each prediction model, 20 runs were conducted to take into account the influence of prediction model parameter randomness and to test the stability of its performance.

2.6. Cascade prediction model

Cascade prediction model is considered where classification at every stage is based on the median value of the corresponding dataset which makes the problem fully balanced throughout the cascade. The first stage of the cascade

model is trained with the whole training set and the classifier predicts if the measurement is above or below the median of the training set. In the training set of every following stage, the measurements that are below 25th percentile of *E. coli* value are removed from the training set, resulting in increased median *E. coli* value of the dataset. 25th percentile and median value were investigated as data reduction limit, however since the reduction of training set size reduces model accuracy, 25th percentile value is chosen as a good measure. This cascading strategy produces overlapping of training data in several stages, which enables the gradual transition towards greater median values and also a gradual reduction in training set size. Flowchart of the proposed methodology can be observed in Figure 3.

With the proposed methodology, the model gradually filters measurements that have low *E. coli* value. If the first stage fails to classify measurement as excellent quality, as a consequence of the gradual transition towards greater *E. coli* values, said measurement can be successfully discarded at the subsequent stage. Additionally, different features' importance is expected, depending on *E. coli* value since the cascade model allows feature weights adjustment at each stage. To improve model reliability, a threshold value for the probability of prediction is introduced. Measurements are considered as excellent quality only if prediction model certainty regarding E.Coli value being below median value is greater than the chosen threshold percentage. Different RF model parameters were investigated and it was observed that change in model parameters greatly influences the efficiency of the threshold

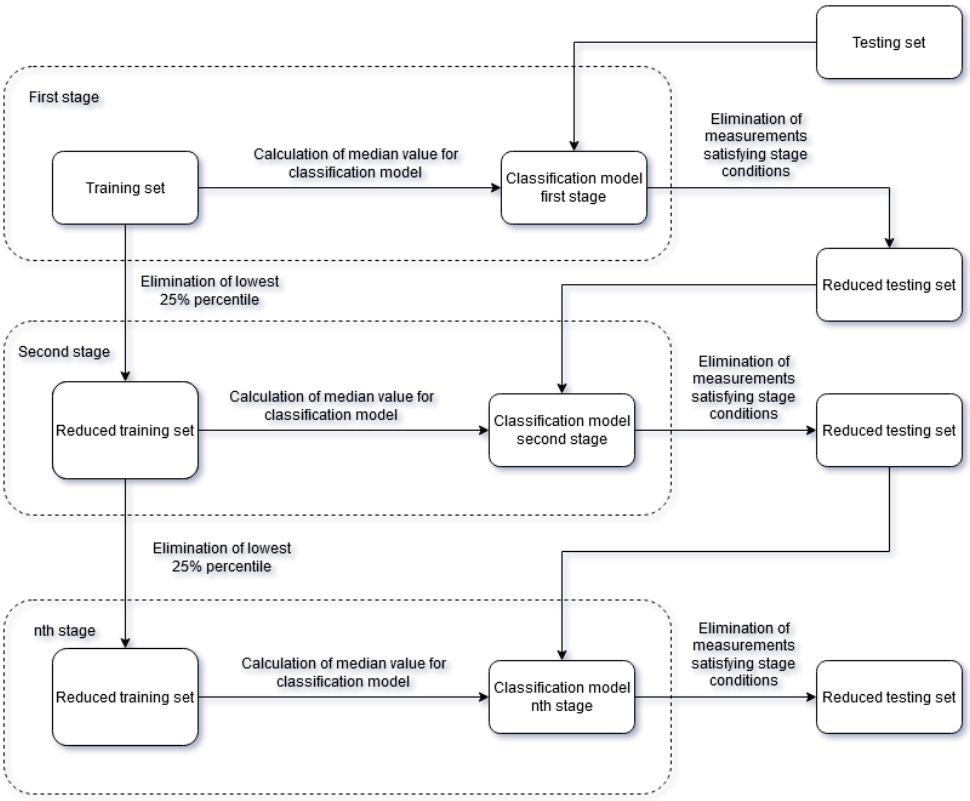

Figure 3: Flowchart of proposed cascade model.

approach. This is because model parameters evaluation based on classification prediction accuracy only considers if classification is true or false, and does not consider model certainty regarding prediction, which is the basis for threshold approach utilisation. Ultimately, the best performance was obtained for 800 estimators, a maximum depth of 10, and the minimum number of samples required to split an internal node equal to 6. These parameters are used for all cascade models and for all stages.

Although each stage of the cascade produces imperfect predictions on

whether *E. coli* concentration is above or below the median value, these wrong predictions are not problematic for the cascade model as a whole, since ultimately what matters is whether the measurement value is above or below the chosen quality limit. Cascade model was investigated for 250 CFU/100 mL limit, which is the EU limit for excellent water quality. This results in 63 measurements (5.5% of all measurements) above the chosen limit, with datasets still having considerable bias.

3. Results

3.1. Random forest classifier - all measuring points

Results for the first group of prediction models with all measurement data for different testing-training data splits and for different classification limits are presented in Table 2. Considered features are water salinity, water temperature, air temperature, GHI, and cumulative GHI for the previous 4 hours which were meteorological data measured during the measurement process with the addition of solar irradiance which is known to positively affect *E. coli* decay (Whitman et al., 2004; Berney et al., 2006; Maraccini et al., 2016). It can be observed that for the EU limit for excellent water quality (250 CFU/100 mL) all models have only 20% of measurements above the given limit correctly classified. That is expected since the number of measurements with reduced water quality is considerably smaller than the number of measurements with excellent quality, thus the prediction model's bias towards excellent quality class yields high model accuracy. When the

national limit for excellent water quality (150 CFU/100 mL) was used, an increased true positive rate can be observed, albeit it is still very low due to a small number of measurements above the chosen limit. Different classification limits were investigated with the addition of median value to create a balanced split between classes. Results are presented in the Figure 4. It was observed that with a lower classification limit problem becomes more balanced and model accuracy increases while model accuracy and true positive rate become similar in proximity to the median limit. To take into account specifics of each training set, it is decided not to consider fixed classification limits, instead, a median value is chosen which always provides a balanced problem through all stages of the cascade model.

Table 2: Prediction model accuracy and true positive rate (TP) for different classification limits of excellent water quality (given in rows) and for different testing sets with indicated number of measurements above considered limit in testing set (given in columns). Results are the average of 20 runs.

	Uniform split		Temporal split		Spatial split	
	Set 1	Set 2	2019	2020	KBW	KW
Number of testing measurements	226	226	100	95	144	151
EU (250 CFU/100 mL)						
Above limit	12	12	16	0	3	12
Model accuracy	94%	95%	83%	/	97%	91%
TP	16%	15%	0%	/	0%	15%
CRO (150 CFU/100 mL)						
Above limit	23	23	26	3	7	25
Accuracy	89%	90%	78%	88%	95%	82%
TP	26%	28%	24%	25%	25%	15%

For each dataset and different classification limits, the analysis of feature

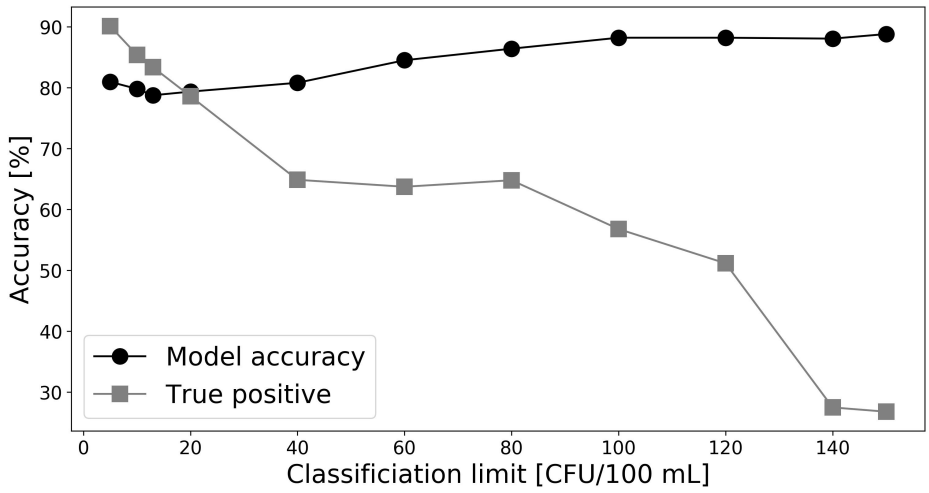

Figure 4: Influence of classification limit on prediction model accuracy and true positive rate for Set 1.

importance was conducted. A similar trend was observed for all considered datasets, thus results are reported only for Set1 (Table 3). It was observed that for both limits, salinity has the greatest importance, followed by GHI, water temperature air temperature, and cumulative GHI for the last 4 hours. Investigation of other classification limits indicated that the prediction model is more uncertain about its decision about feature importance when a higher classification limit is chosen, which can also be observed here with greater standard deviation for features for EU (250 CFU/100 mL) limit for excellent water quality.

3.2. Random forest classifier - separated models

Further analysis was conducted for two separated models, one with a group of measuring points with higher average *E. coli* value and the second

Table 3: Prediction model feature importance for Set1 for different classification limits of excellent water quality. Results are average of 20 runs and numbers in brackets indicate standard deviation.

	Limit	
Features	EU (250 CFU/100 mL)	CRO (150 CFU/100 mL)
Salinity	26% (0.8%)	29% (0.6%)
GHI	23% (1.1%)	21% (0.7%)
Cumulative GHI	16% (0.6%)	15% (0.6%)
Water temp.	18% (0.9%)	19% (0.6%)
Air temp.	17% (0.8%)	15% (0.5%)

for a group of measuring points with smaller average *E. coli* value. All models were tested on Set1 with uniformly distributed *E. coli* measurements. In order to account for the influence of the training dataset sizes, the datasets for both the model for low *E. coli* values and the model for all measuring points were reduced so as to be of the approximately same size as the model with high *E. coli* values. Each n-th measurement is removed from the dataset, so the uniform distribution of *E. coli* measurements is maintained. The results are presented in Table 4. It can be observed that the classification limit has the greatest influence on prediction model accuracy. Grouping of similar measurement locations does not contribute to better prediction accuracy, since for all combinations of grouping and number of inputs, the true positive rate is still below 30%. The greatest true positive rate (82%) is achieved for the prediction model with all measurement points and for all available data when the median value is taken as a classification limit. Although this behaviour should be investigated for other training-testing splits, in the subsequent analysis of the cascade model, a single prediction

model with all measurement points will be adopted for each stage, as it is evident that the number of measurements, when compared to segregated approach, contributes more to the overall accuracy.

Table 4: Prediction model average accuracy for 20 runs for the three considered models with approximately equal number of inputs and with unreduced dataset. Features considered are salinity, water and air temperature, GHI and cumulative GHI for 4 hours.

Low *E. coli*	394 train, 99 test		526 train, 131 test	
	CRO	Median (6.5)	CRO	Median (6.5)
Model acc	95%	70%	94%	71%
TP	16%	61%	13%	58%
High *E. coli*	381 train, 95 test			
	CRO	Median (38)		
Model acc	83%	75%		
TP	19%	73%		
All measuring points	383 train, 96 test		907 train, 226 test	
	CRO	Median (14)	CRO	Median (13)
Model acc	79%	79%	89%	78%
TP	26%	78%	27%	82%

3.3. Feature analysis

Further investigation was conducted for rainfall features for the prediction model with all measurement points and median as classification limit. It can be observed from Table 5 that the inclusion of considered rainfall features does not contribute to a significant change in prediction model accuracy. It is interesting to observe that the cumulative sum that accounts for rainfall from the 4th to 7th day prior has lower importance than the cumulative sum that accounts for rainfall from the 7th to 14th day prior. Due to having the lowest feature importance, precipitation features were not included in the

cascade model testing. However, since it does not reduce model accuracy either, further study should be conducted where inclusion of precipitation features at higher cascade model stages, with greater median values, could be more beneficial.

Table 5: Feature importance for prediction model with all measurement points with median value as classification limit. Results are average of 20 runs.

Features	Group1	Group2	Group3	Group4
Salinity	38%	41%	40%	43%
Water temperature	11%	12%	11%	12.5%
Air temperature	10%	11%	10%	11.5%
Cumulative GHI	11%	13%	12%	14%
GHI	15%	17%	17%	19%
Rainfall 4-7 days	6%	7%	/	/
Rainfall 7-14 days	9%	/	10%	/
Model acc.	81%	80%	81%	80%
TP	83%	83%	83%	83%

3.4. Cascade model results

The cascade model was first tested on Set1 in order to calibrate the stage parameters. The overview of stage metrics for Set1 can be observed in Table 6. Gradual reduction of training inputs and increase of median value through stages can be observed. Six stages were chosen since it was decided that with further stages the size of the training dataset would be too low for prediction model training.

The cascade model was run 50 times, where average feature importance through stages is presented in Table 7. It can be observed that at the first stage water salinity has the greatest importance, although as prediction mod-

Table 6: Stages parameters for Cascade model for training Set 1.

Stage	Number of training inputs	*E. coli* value median	25% percentile	training range
1	907	13	3	all
2	690	30	9	$\geq$3
3	520	55	20	$\geq$9
4	397	80	42	$\geq$20
5	298	120	70	$\geq$42
6	231	130	92	$\geq$70

els are trained on measurements with greater *E. coli* value, salinity importance decreases, where other features are given greater importance. This indicates that the introduction of other features at higher level stages could be beneficial for the prediction model.

Table 7: Feature importance for cascade model for training Set 1.

Features	Stage					
	1	2	3	4	5	6
Salinity	43%	35%	31%	29%	22%	22%
GHI	18%	18%	20%	23%	25%	27%
Cumulative GHI	14%	18%	18%	18%	17%	17%
Water temperature	12%	14%	16%	18%	19%	20%
Air temperature	12%	14%	15%	13%	17%	15%

The influence of threshold value on model accuracy results can be observed in Table 8. The same threshold value is set for all stages. Greater number of measurements are predicted as excellent quality with a lower threshold value, albeit with a greater percentage of wrong predictions. For most datasets, the threshold value of 80% assures there is no elimination of days with *E. coli* value greater than 250 CFU/100 mL, with the exception of

a single measurement for the KW dataset which is still incorrectly predicted even with a threshold of 85%. This measurement could be an outlier or additional cascade model improvement could eliminate this wrong prediction. Thus, for the purpose of further study, the threshold value of 80% is adopted.

Table 8: Influence of threshold value on cascade model accuracy for different datasets. Numbers in brackets next to datasets indicate total number of measurements in testing set and number of *E. coli* measurements above EU quality limit. Presented results are average of 50 runs.

Set1 (226, 15)	Threshold			
Excellent quality prediction	85%	80%	75%	70%
True positive	72 (34%)	91 (43%)	106 (50%)	125 (60%)
False negative	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Set2 (226, 15)				
True positive	65 (31%)	96 (45%)	117 (55%)	140 (66%)
False negative	0 (0%)	0 (0%)	0.28 (2%)	1 (7%)
KBW (144, 3)				
True positive	82 (58%)	84 (59%)	96 (68%)	105 (74%)
False negative	0 (0%)	0 (0%)	0 (0%)	0 (0%)
KW (151,12)				
True positive	34 (24%)	55 (39%)	71 (51%)	84 (61%)
False negative	1 (8%)	1 (8%)	1.7 (14%)	3 (25%)
2019 (100,16)				
True positive	4 (5%)	11 (13%)	27 (32%)	41 (48%)
False negative	0 (0%)	0 (0%)	1 (6%)	2.3 (14%)
2020 (95, 0)				
True positive	34 (36%)	50 (52%)	59 (62%)	70 (74%)
False negative	0 (0%)	0 (0%)	0 (0%)	0 (0%)

To further enhance the proposed model, a combination of threshold values for different stages was analyzed. Since in the first few stages the median value is considerably low, it is reasonable to assume that a smaller threshold value in those stages could still provide good results. For the first stage

threshold value of 65% is considered, for the second stage 70%, for the third stage 75%, and for all other stages 80%. Secondly, due to the overlapping of training data through stages, an additional check is introduced: if both in current and previous stage prediction model certainty is above the chosen threshold, which is chosen to be lower than the threshold for the current stage, then measurement can also be classified as below limit. This can be understood as that two weak certainties at subsequent stages can be considered as one strong certainty in the current stage. The considered threshold values were 70% for the second and third stage and 75% for the remaining stages. Additionally, the influence of different features throughout the stages was also considered. Due to its small feature importance in the first three stages, the air temperature was removed as a feature and then introduced only in the last three stages. Combinations of these methods were also investigated.

Overview of the obtained results can be seen in Table 9. It can be observed that all proposed methods increase the number of excellent quality predictions, where the combination of all three methods further increases that number. It must be noted that only one combination for each method was presented, where further investigation of more combinations could provide better results, e.g. different features, different values of increasing threshold value, etc.

Proposed improvement of the cascade model with all three adjustments was tested on different datasets and results are presented in Table 10. It can be observed that the number of correct predictions of excellent quality is

Table 9: Influence of stage adjustment on cascade model for Set1. Presented results are average of 50 runs with no wrong predictions of excellent water quality.

	Remaining measurements	Right prediction (%)
Feature change	136	43%
Additional deletion	125	48%
Increasing threshold (65-80%)	116	52%
All	113	54%

greatly influenced by the testing set. Also, for some datasets wrong predictions are also observed, indicating that more rigorous adjustment of proposed cascade model improvements should be conducted.

Table 10: Cascade model results for prediction of excellent water quality with all 3 methods for different datasets.

Dataset	Number of measurements	Measurements with reduced quality	True positive	False negative
2019	100	16	27 (32%)	0.02 (0.13%)
2020	95	0	61 (65%)	0 (0%)
KBW	144	3	71 (51%)	0 (0%)
KW	151	3	71 (51%)	0 (0%)
Set1	226	15	113 (54%)	0 (0%)
Set2	226	15	118 (56%)	1.24 (8%)

4. Discussion

4.1. Prediction modeling

From the conducted analysis it can be concluded that a single prediction model for prediction of *E. coli* value being above or under limits 150 or 250 CFU/100 mL does not provide satisfactory results due to considerable data bias. The prediction model tends to classify all measurements as excellent

quality since they make a vast majority of the measurements. Thus, a cascade model is introduced which is shown to work well as a filter of measurements with excellent quality, without removing measurements above the chosen limit. The proposed cascade model has balanced datasets at each stage since the median value is considered as a classification limit at each stage. However, different limits can also be explored to possibly further improve model accuracy. It must be noted that in general, measurements considered in this paper have overall excellent water quality, however, the presented methodology can be used for different pollution values since the cascade model is constructed to adapt to the provided data.

It was observed that the salinity feature has the greatest importance in single prediction models, and also for several first stages of the cascade model. However, in further stages, it was observed that salinity value is not as important. It indicates that the salinity feature has the greatest importance for predicting measurements with low *E. coli* value. Since the majority of measurements for the single prediction model have low *E. coli* value, a strong weight is put on the salinity value which is beneficial for the majority of data but is not beneficial for the right prediction of high *E. coli* measurements, which are in fact most important. Thus, the cascade model enables adjustment of feature importance through stages for different levels of pollution. Additionally, some features that could be important for the right prediction of higher *E. coli* values would decrease accuracy for low *E. coli* values, thus different features can be included at different stages. In

order to maximize the capabilities of the cascade model approach, precise feature engineering should be conducted.

It must be noted that RF model parameters were the same for all stages and proposed improvements for the cascade model were investigated only for one combination of parameters. Further investigation of model parameters through stages and further study of different combinations of proposed improvements should be conducted to further increase cascade model efficiency. Currently, both cascade and single prediction models are constructed with measurements from multiple measurement points due to the small amount of data. However, ultimately separated prediction models could be constructed to establish a relation between *E. coli* value and specific processes for that measurement point.

4.2. Data analysis

From in-depth analysis of measurement data, it was observed that sources of groundwater are related with the greater value of *E. coli*. This was also corroborated by the higher importance of salinity in the first couple of stages of the cascade model. High salinity value corresponds to summer months with longer dry periods, where *E. coli* values are very low (as are classification limits for the first several stages), where the prediction model heavily relies on that information to consider a measurement as excellent quality. As salinity value decreases, it most often corresponds to spring months where the sea is still influenced by longer periods of rainfall, or alternatively to measurements

influenced by coastal springs. For both of these occurrences, it is important to consider solar irradiation and both water and air temperature, which is supported with more evenly distributed feature importance at higher stages.

Groundwater springs are of great interest when they can be used as a source of drinking water. Since coastal springs that are of interest in this study are brackish springs, smaller and with seasonal character there is no study of their characteristics in the literature. Additional in-field investigation of 3M location was conducted during May 2021 after a longer period of rain where a considerable number of submerged spring sources were observed (Figure 5). Although the occasional presence of springs is well known since they can be experienced as cold areas during swimming, they were not previously considered as being of strong importance and were thought to dry up during summer. However, observation of these springs is mentioned in Stražičić (1999), where two big springs which are active throughout the whole year are said to exist at the 3M location. Furthermore, another three springs are mentioned, where the eastern spring, which is closest to the bathing location, is active yearlong and the other two usually become inactive only by the end of summer. In KE location one strong spring is mentioned with the addition of the number of smaller springs throughout the coast.

Additionally, in the immediate vicinity of the beach 3M is greater spring Cerovice (Figure 6) which is located inside shipyard "3. Maj". It consists of several springs that are active throughout the whole year, but since microbiological pollution was observed that water is only used as technical water.

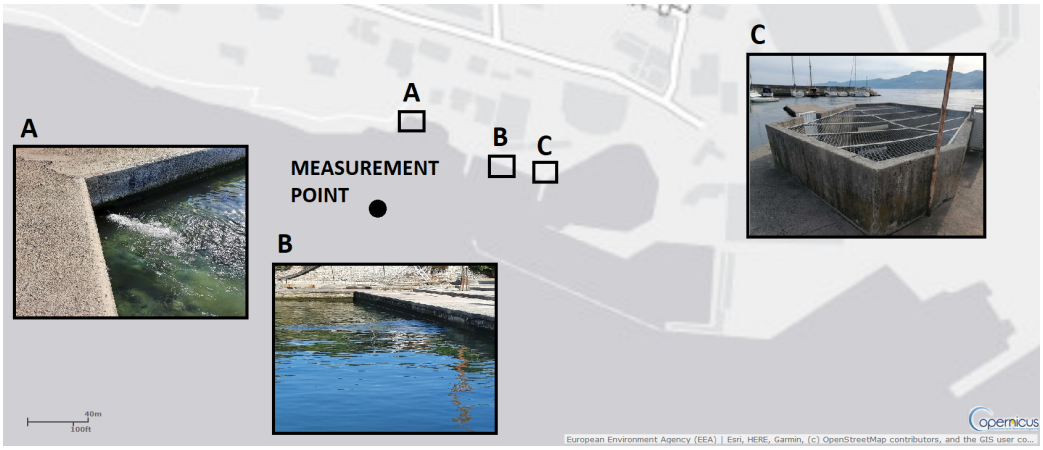

Figure 5: Detail of 3M beach location with indicated some of the springs locations. Location C is collector of multiple springs.

This could explain reduced water quality for measurement locations 3M and KE, which are specific since they have a considerable amount of springs active yearlong, probably with water quality similar to spring Cerovice. Reduced water quality in KW location could be due to the influence of KE springs since wind and sea currents can cause transport of contamination. Additionally, in the study by Biondić et al. (2009) risk assesment of underground water was conducted, where it was indicated that landfills in the hinterland of city Rijeka can influence coastal springs in the area considered in this study (Figure 6).

These observations indicate that current data, comprising both from regular measurements taken from the boat (approximately 20 m from the coastline) and additional measurements taken from the coast, could give considerably different results under the same meteorological conditions if additional sampling is conducted in the immediate vicinity of the spring location. Both

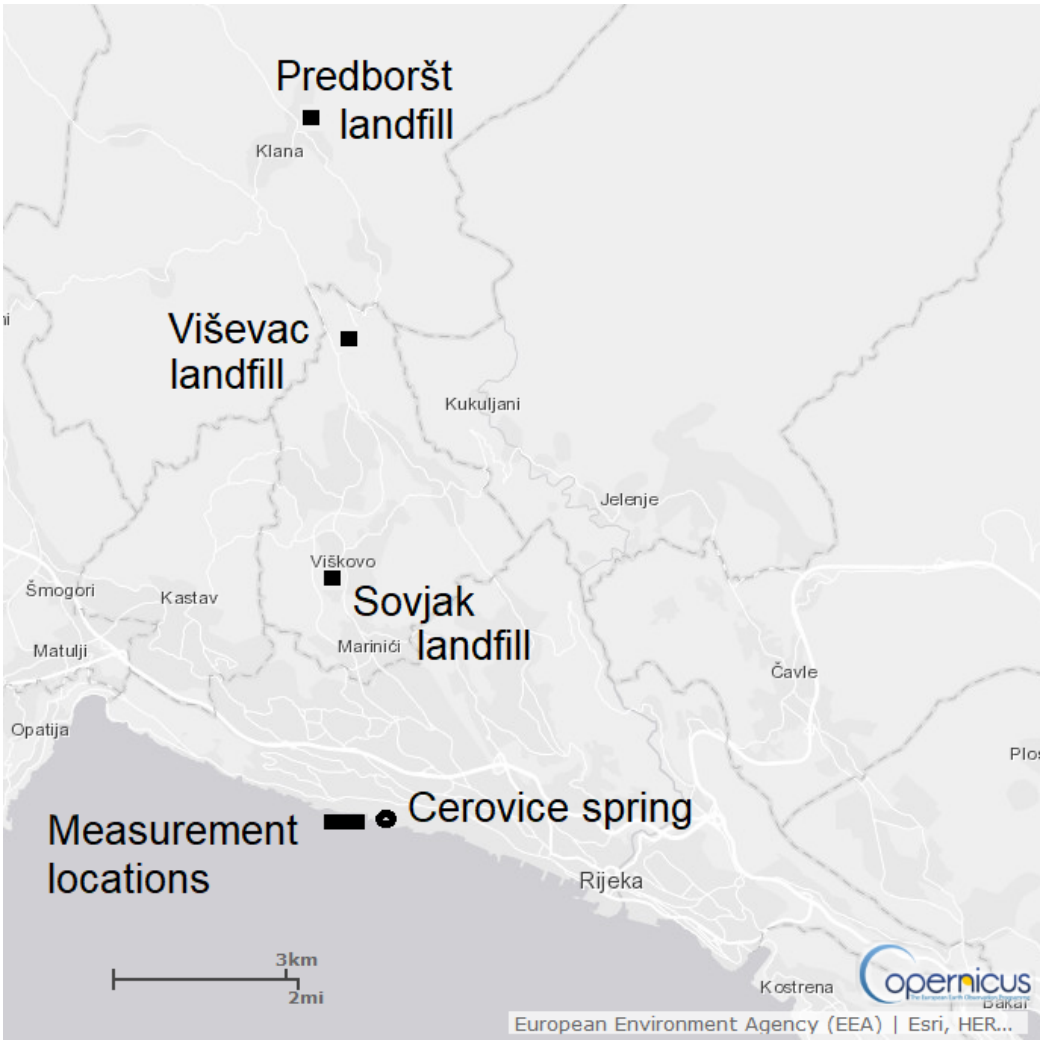

Figure 6: Locations of spring Cerovice and landfills in the hinterland of measurement locations.

types of measurements were included to increase the teaching dataset which should lead to improved model accuracy, however with these new findings, this could in fact reduce prediction model accuracy. Thus, several directions of further research are suggested.

4.3. Further research

First, analysis of the coastal area should be conducted in order to investigate the number and locations of possible coastal springs. It is known that some sources are active only for some periods, where some sources are active almost all year. Thus, an investigation of the source activation period should be conducted. Observation of these sources presented in Stražičić (1999) only notes their existence and locations while activation periods and discharge values should be also known so as to assess their influence on bathing locations. Also, these observations should be revised and adjusted, due to changes in city infrastructure (such as drainage and sewage systems reconstructions) conducted in the period of the last 20 years or more. Additionally, every beach location has its specific geomorphology, where the number and size of springs are different, as well as its openness to the sea. In this study, 3M location was further inspected (Figure 5) which is a closed, small port, and as such under a strong influence of the observed springs. Other locations, which are less enclosed, are expected to show a smaller influence of springs, however, this premise should be further investigated.

The second direction of future research should investigate rainfall influence and watershed area more extensively. Different rainfall correlations with increased microbial pollution are found through literature, e.g. previous 24 hours (Mallin et al., 2001), 48 hours (Kelsey et al., 2004), 7 days (Lipp et al., 2001), 5.5-9.5 days (de Souza et al., 2018b) and even 30 days (Pandey et al., 2012). This is a strong indication that geological specifics

must be taken into consideration. Precipitation considered in this work was only from a single location for the city center of Rijeka, which can influence short-term water influx to the sea. However, during summer periods it can be expected that dry ground absorbs a great amount of rainfall. This could explain why there was no considerable correlation with cumulative sums of rainfall from several previous days in the city center. Additionally, karst groundwater is greatly influenced by precipitation from the whole watershed area. For example, tracer tests were conducted in Biondić et al. (2005) where the tracer was injected in V. Snežnik (Slovenia) where underground water connection with Kvarner Bay area was identified, including measurement locations considered in this study, indicating transboundary characteristics of the investigated area. Thus, further research should include an investigation of correlations between coastal springs' activation and wider regional area's precipitation, as these distant rainfalls are expected to possibly be more influential for coastal spring activity than local rainfall. It is also important to mention that a boundary between two watersheds is passing through the center of the city of Rijeka, thus bathing locations on the east side of the city are expected to show a different behaviour, and different locations for rainfall measurements should be considered.

Ultimately, the measurement process of *E. coli* could be improved on the grounds of these new observations. If a stronger influence of groundwater sources is observed for some bathing locations, and if those sources are observed to often have reduced water quality, a unique measurement method-

ology should be established for these locations, where multiple locations, contrary to the current single measurement location, should be investigated to give a better description of the considered bathing location. Additionally, in Rukavina (2007) the bathing locations in the east part of Rijeka were analysed, and springs were found to have unusually and considerably greater value of *E. coli* then values observed at the monitoring location. Follow-up investigation led to the identification of one business located approximately 2 km upstream, with inappropriate wastewater connection as a source of pollution which was connected with observed springs. This indicates that springs are a great vulnerability of bathing locations, especially if they have a greater outflow and are influenced by greater watershed area where periodic monitoring of such springs should also be considered.

5. Conclusion

In the presented paper, in-depth data analysis of *E. coli* measurements and related data was conducted and a predictive machine learning modeling strategy was proposed. Due to a very small amount of measurements with reduced water quality in the database, it was shown that a single prediction model has reduced prediction accuracy due to bias toward classifying all measurements as excellent quality. Thus, a cascade model approach is proposed which classifies measurement as excellent quality only if it is highly certain regarding its decision. Other measurements remain suspect, therefore the proposed method can be considered as a filter method, which can be fur-

ther improved to be used as a standalone model. The following observations regarding the prediction model can be made:

- Grouping of measurement points does not provide improvement in prediction model accuracy since an increased number of inputs is more beneficial.
- Due to bias of input data it is difficult for a single prediction model to confidently predict occurrences of subpar bathing water quality.
- Cascade model can provide a predictive data filter in which excellent water quality can be predicted with high accuracy, based on meteorological data, solar irradiation, and seawater salinity.
- Due to the high flexibility of cascade model, multiple strategies of its improvement could possibly lead to it being eventually used as a standalone prediction model.
- Ultimately, a separate prediction model for each measurement point could be constructed, to capture the uniqueness of each beach, which is especially important in the karst type of terrain.

Conducted study indicates a strong need for an interdisciplinary approach for the given problem. The karst type of soil with its specific underground landscape absorbs rainfall from a large watershed area, indicating not only rainfall period but also rainfall measurement locations are important. Additionally, *E. coli* measurements could be conducted for the number of known

springs to indicate which springs contribute to the coastal seawater pollution and in which amount. Regarding *E. coli* measurements, the following is observed:

- Submerged groundwater sources, active yearlong, are observed which could correlate with greater *E. coli* value.
- Number of groundwater sources and their intensity should be investigated at the considered bathing locations.
- New sampling methodology should be established to take into consideration the mixing of microbiologically contaminated submerged spring water and seawater.

6. Acknowledgements

This research article is a part of the project *Computational fluid flow, flooding, and pollution propagation modeling in rivers and coastal marine waters - KLIMOD* (grant no. KK.05.1.1.020017), and is funded by the Ministry of Environment and Energy of the Republic of Croatia and the European structural and investment funds. Also, the authors acknowledge the funding of the University of Rijeka through the project *Razvoj hibridnog 2D/3D modela za učinkovito modeliranje strujanja u rijekama, jezerima i morima.* Furthermore, authors acknowledge the support of the Center of Advanced Computing and Modelling at the University of Rijeka for providing computing resources.

References

Agency, E.E., 2009. Bathing water profiles: Best practice and guidance. URL: https://ec.europa.eu/environment/water/water-bathing/pdf/profiles_dec_2009.pdf. accessed: 2021-03-06.

Berney, M., Weilenmann, H.U., Simonetti, A., Egli, T., 2006. Efficacy of solar disinfection of escherichia coli, shigella flexneri, salmonella typhimurium and vibrio cholerae. Journal of applied microbiology 101, 828–836.

Biondić, R., Biondić, B., Rubinić, J., Meaški, H., Kapelj, S., Tepeš, P., 2009. Ocjena stanja i rizika cjelina podzemnih voda na krškom području u republici hrvatskoj. Završno izvješće. Arhiv Geotehničkog fakulteta Sveučilišta u Zagrebu, Varaždin .

Biondić, R., Prestor, J., Kapelj, S., Dolić, S., 2005. Slovenski snežnik-visoka zona sliva izvora grada rijeke. Knjiga sažetaka-3. Hrvatski geološki kongres , 179–180.

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

Buckerfield, S.J., Quilliam, R.S., Waldron, S., Naylor, L.A., Li, S., Oliver, D.M., 2019. Rainfall-driven e. coli transfer to the stream-conduit network observed through increasing spatial scales in mixed land-use paddy farming karst terrain. Water research X 5, 100038.

Chahouri, A., El Ouahmani, N., El Azzaoui, A., Yacoubi, B., Banaoui, A.,

Moukrim, A., 2021. Combined assessment of bacteriological and environmental indicators of fecal contamination in agadir bay ecosystems (southwest morocco). International Journal of Environmental Science and Technology , 1–14.

Cho, K.H., Cha, S.M., Kang, J.H., Lee, S.W., Park, Y., Kim, J.W., Kim, J.H., 2010. Meteorological effects on the levels of fecal indicator bacteria in an urban stream: a modeling approach. Water research 44, 2189–2202.

Davis, R.K., Hamilton, S., Brahana, J.V., 2005. Escherichia coli survival in mantled karst springs and streams, northwest arkansas ozarks, usa 1. JAWRA Journal of the American Water Resources Association 41, 1279–1287.

Directive, G.o.t.R.o.C., 2008. Regulation on sea bathing water quality, government of the republic of croatia. Official Gazette No. 73.

Dwight, R.H., Caplan, J.S., Brinks, M.V., Catlin, S.N., Buescher, G., Semenza, J.C., 2011. Influence of variable precipitation on coastal water quality in southern california. Water Environment Research 83, 2121–2130.

EC, 2006. European council directive 2006/7/ec of 15 february 2006 concerning the management of bathing water quality and repealing directive 76/160/eec. Official Journal of the European Union, L64/37.

Ekklesia, E., Shanahan, P., Chua, L.H., Eikaas, H., 2015. Temporal variation of faecal indicator bacteria in tropical urban storm drains. Water research 68, 171–181.

European Union, Copernicus Land Monitoring Service 2018, E.E.A.E., 2018. Copernicus land monitoring service. URL: https://land.copernicus.eu/.

Grbčić, L., Družeta, S., Mauša, G., Lipić, T., Lušić, D.V., Alvir, M., Lučin, I., Sikirica, A., Davidović, D., Travaš, V., et al., 2021. Coastal water quality prediction based on machine learning with feature interpretation and spatio-temporal analysis. arXiv preprint arXiv:2107.03230 .

He, L.M.L., He, Z.L., 2008. Water quality prediction of marine recreational beaches receiving watershed baseflow and stormwater runoff in southern california, usa. Water research 42, 2563–2573.

He, Y., He, Y., Sen, B., Li, H., Li, J., Zhang, Y., Zhang, J., Jiang, S.C., Wang, G., 2019. Storm runoff differentially influences the nutrient concentrations and microbial contamination at two distinct beaches in northern china. Science of The Total Environment 663, 400–407.

Heinz, B., Birk, S., Liedl, R., Geyer, T., Straub, K., Andresen, J., Bester, K., Kappler, A., 2009. Water quality deterioration at a karst spring (gallusquelle, germany) due to combined sewer overflow: evidence of bacterial and micro-pollutant contamination. Environmental Geology 57, 797–808.

Jozić, S., Lušić, D.V., Ordulj, M., Frlan, E., Cenov, A., Diković, S., Kauzlarić, V., Đurković, L.F., Totić, J.S., Ivšinović, D., et al., 2018. Performance characteristics of the temperature-modified iso 9308-1 method for the enumeration of escherichia coli in marine and inland bathing waters. Marine pollution bulletin 135, 150–158.

Jozić, S., Vukić Lušić, D., 2018. Report on validation of temperature modified iso 9308-1: 2014 method for the enumeration of escherichia coli in bathing water samples in croatia. Institute of Oceanography and Fisheries: Croatia, Balkan .

Kelsey, H., Porter, D., Scott, G., Neet, M., White, D., 2004. Using geographic information systems and regression analysis to evaluate relationships between land use and fecal coliform bacterial pollution. Journal of experimental marine biology and ecology 298, 197–209.

Knierim, K.J., Hays, P.D., Bowman, D., 2015. Quantifying the variability in escherichia coli (e. coli) throughout storm events at a karst spring in northwestern arkansas, united states. Environmental earth sciences 74, 4607–4623.

Kogovšek, J., Petrič, M., 2013. Increase of vulnerability of karst aquifers due to leakage from landfills. Environmental earth sciences 70, 901–912.

Kucuksezgin, F., Gonul, L.T., Pazi, I., Kacar, A., 2019. Assessment of seasonal and spatial variation of surface water quality: Recognition of en-

vironmental variables and fecal indicator bacteria of the coastal zones of izmir bay, eastern aegean. Regional Studies in Marine Science 28, 100554.

Laroche, E., Petit, F., Fournier, M., Pawlak, B., 2010. Transport of antibiotic-resistant escherichia coli in a public rural karst water supply. Journal of Hydrology 392, 12–21.

Lipp, E.K., Kurz, R., Vincent, R., Rodriguez-Palacios, C., Farrah, S.R., Rose, J.B., 2001. The effects of seasonal variability and weather on microbial fecal pollution and enteric pathogens in a subtropical estuary. Estuaries 24, 266–276.

Mallin, M.A., Ensign, S.H., McIver, M.R., Shank, G.C., Fowler, P.K., 2001. Demographic, landscape, and meteorological factors controlling the microbial pollution of coastal waters, in: The Ecology and Etiology of Newly Emerging Marine Diseases. Springer, pp. 185–193.

Mance, D., Mance, D., Vukić Lušić, D., 2018. Environmental isotope\(ˆ 18 o\) in coastal karst spring waters as a possible predictor of marine microbial pollution. Acta Adriatica: international journal of Marine Sciences 59, 3–15.

Maraccini, P.A., Mattioli, M.C.M., Sassoubre, L.M., Cao, Y., Griffith, J.F., Ervin, J.S., Van De Werfhorst, L.C., Boehm, A.B., 2016. Solar inactivation of enterococci and escherichia coli in natural waters: effects of water absorbance and depth. Environmental science & technology 50, 5068–5076.

Pandey, P.K., Soupir, M.L., Haddad, M., Rothwell, J.J., 2012. Assessing the impacts of watershed indexes and precipitation on spatial in-stream e. coli concentrations. Ecological indicators 23, 641–652.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: Machine learning in python. Journal of machine learning research 12, 2825–2830.

Pikelj, K., Juračić, M., 2013. Eastern adriatic coast (eac): geomorphology and coastal vulnerability of a karstic coast. Journal of coastal research 29, 944–957.

Reed, T.M., Fryar, A.E., Brion, G.M., Ward, J.W., 2011. Differences in pathogen indicators between proximal urban and rural karst springs, central kentucky, usa. Environmental Earth Sciences 64, 47–55.

Rukavina, T., 2007. Važnost monitoringa mikrobioloških indikatora u vodama priobalnih izvora. Zbornik radova 4. hrvatske konferencije o vodama Hrvatske vode i Europska Unija-izazovi i mogućnosti , 217–222.

Solcast, 2021. Solar irradiance data. URL: `https://solcast.com`. accessed: 2020-12-01.

Solo-Gabriele, H.M., Harwood, V.J., Kay, D., Fujioka, R.S., Sadowsky, M.J., Whitman, R.L., Wither, A., Caniça, M., Da Fonseca, R.C., Duarte, A.,

et al., 2016. Beach sand and the potential for infectious disease transmission: observations and recommendations. Journal of the Marine Biological Association of the United Kingdom 96, 101–120.

de Souza, R.V., Campos, C.J., Garbossa, L.H., Seiffert, W.Q., 2018a. Developing, cross-validating and applying regression models to predict the concentrations of faecal indicator organisms in coastal waters under different environmental scenarios. Science of The Total Environment 630, 20–31.

de Souza, R.V., de Campos, C.J.A., Garbossa, L.H.P., de Novaes Vianna, L.F., Seiffert, W.Q., 2018b. Optimising statistical models to predict faecal pollution in coastal areas based on geographic and meteorological parameters. Marine pollution bulletin 129, 284–292.

Stange, C., Tiehm, A., 2020. Occurrence of antibiotic resistance genes and microbial source tracking markers in the water of a karst spring in germany. Science of The Total Environment 742, 140529.

Stražičić, N., 1999. Riječki izvori i vodotoci. Izdavački centar Rijeka, Rijeka.

Verga, R.N., Tolosano, J.A., Cazzaniga, N.J., Gil, D.G., 2020. Assessment of seawater quality and bacteriological pollution of rocky shores in the central coast of san jorge gulf (patagonia, argentina). Marine pollution bulletin 150, 110749.

Viau, E.J., Goodwin, K.D., Yamahara, K.M., Layton, B.A., Sassoubre, L.M., Burns, S.L., Tong, H.I., Wong, S.H., Lu, Y., Boehm, A.B., 2011. Bacterial pathogens in hawaiian coastal streams—associations with fecal indicators, land cover, and water quality. Water research 45, 3279–3290.

Vukić Lušić, D., Kranjčević, L., Maćešić, S., Lušić, D., Jozić, S., Linšak, Ž., Bilajac, L., Grbčić, L., Bilajac, N., 2017. Temporal variations analyses and predictive modeling of microbiological seawater quality. Water research 119, 160–170.

Weiskerger, C.J., Phanikumar, M.S., 2020. Numerical modeling of microbial fate and transport in natural waters: Review and implications for normal and extreme storm events. Water 12, 1876.

Whitman, R.L., Nevers, M.B., Korinek, G.C., Byappanahalli, M.N., 2004. Solar and temporal effects on escherichia coli concentration at a lake michigan swimming beach. Applied and environmental microbiology 70, 4276–4285.

Zaninović, K., Gajić-Čapka, M., Perčec Tadić, M., Vučetić, M., Milković, J., Bajić, A., Cindrić, K., Cvitan, L., Katušin, Z., Kaučić, D., et al., 2008. Klimatski atlas hrvatske. Državni hidrometeorološki zavod, Zagreb 172.