Water Research 177 (2020) 115788

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres

A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning

Tingting Xu*, Giovanni Coco, Martin Neale

School of Environment, Faculty of Science, University of Auckland, New Zealand

ARTICLE INFO

Article history:
Received 9 December 2019
Received in revised form
1 April 2020
Accepted 2 April 2020
Available online 13 April 2020

Keywords:
Water quality
Adaptive synthetic sampling algorithm
Nearest neighbour
Boosting decision tree
Support vector machine
Artificial neural network

ABSTRACT

Predicting recreational water quality is one of the most difficult tasks in water management with major implications for humans and society. Many data-driven models have been used to predict water quality indicators to allow a real time assessment of public health risk. This assessment is most commonly based on Faecal Indicator Bacteria (FIB), with the value of FIB compared with thresholds published in guidelines. However, FIB values usually tend to be unbalanced within water quality datasets, with small proportions of data exceeding guideline thresholds and far larger numbers that do not. This can be a limiting factor in the uptake of model predictions since, even if the overall accuracy is high, the sensitivity of the predictions can be low. To address this issue, this paper proposes an adaptive synthetic sampling algorithm (ADASYN) to generate synthetic above-threshold FIB instances and test the validity of the approach for the prediction of recreational water quality. The models in this paper are based on four machine learning techniques: k-mean nearest neighbour, boosting decision tree, support vector machine, and multi-layer perceptron artificial neural network and are applied to five different locations in Auckland, New Zealand. Aside from support vector machine, all models provide favourable predictions with relatively high sensitivity (around 75%) and overall accuracy (over 90%), indicating that both the compliant and exceedance conditions can be effectively predicted through the use of more sophisticated model training which involves artificial data. Considering the model accuracy and stability, boosting decision trees (BDT) and multi-layer perceptron artificial neural (MLP-ANN) network are the best two models and the multi-layer perceptron is the most efficient with the shortest computation time.

1. Introduction

Robust and reliable information about the water quality at beaches is of vital importance as it is used to inform managers about public health risk from recreation in water (Thoe and Lee, 2013; Stidson et al., 2012; Francy et al., 2013; Wade et al., 2008). Beach water quality is widely monitored using faecal indicator bacteria (FIB), which commonly involves the measurement of enterococci or *Escherichia coli* concentrations in water. Monitoring results are usually assessed for compliance with guidelines, and guideline exceedances trigger advisories or warnings for beach users (de Brauwere et al., 2014; Castelletti et al., 2014; Marshall et al., 1997; King and Richardson, 2003). The laboratory tests for FIB typically take up to 24 h using the most probable method (MPN), however, once the time for sample collection and interpretation and communication of results are included, the test result is not publicly available probably for up to 48 h after sample collection. This results in water quality information that is potentially out of date because water quality can change faster than the analysis time. For this reason, much effort has been invested in trying to predict FIB concentrations in real time, with a range of process and data driven based models utilised for this purpose (e.g. Bae et al., 2010; Abyaneh, 2014; Thoe and Lee, 2013; He and He, 2008).

The vast majority of models used to predict water quality are data-driven, statistical and categorical, and are therefore sensitive to the data used in their development (Thoe et al., 2014; Chen and Liu, 2015; Li et al., 2019). For example, all these models faced the issue of unbalanced datasets, whereby the number of data points that are compliant with guidelines is always much more than the data exceeding the guideline thresholds for FIB. Therefore, the models are usually good at predicting compliant conditions rather exceedance condition. Chandramouli et al. (2007) and Tufail et al.

* Corresponding author.
E-mail address: txu648@aucklanduni.ac.nz (T. Xu).

https://doi.org/10.1016/j.watres.2020.115788
0043-1354/© 2020 Elsevier Ltd. All rights reserved.

(2008) used artificial neural networks (ANN) to predict water quality and found that they slightly outperformed traditional regression models. However, they only considered the overall accuracy which can lead to a significant bias due to the original dataset which contains a disproportionately large proportion of compliant data points (i.e., when the FIB is below the guideline threshold). In 2010, Kazemi Yazdi and Scholz, assessed water quality based on surface runoff through an ANN model and compared the ANN with a multi-linear regression model (MLR). The ANN outperformed MLR in predicting the runoff treatment since it has a higher ability to capture the nonlinear relationship between water quality and microbial factors. They focused on the exploration of the relation between the quality indicator and the impact factor but failed to provide details on the model validation aspect. Qin et al. (2012) introduced an advanced Boosting-tree based machine learning model but did not compare their results to other models. Thoe et al. (2014) applied three regression models as well as classification trees (CT) and ANN. They assessed model results in terms of overall accuracy, but also sensitivity and specificity, and concluded that CT and ANN performed better than regression models. Here the overall accuracy refers to the model successfully predicting below and above threshold data, sensitivity assesses the model's ability to predict only exceeding data, and specificity assesses the prediction of compliant data. Even in this case, sensitivity was still low, probably as a result of the bias to the much more non-exceeding data in the training dataset (Bedri et al., 2016; Wang et al., 2016; Shaw et al., 2017). Besides these models, the capability of multi-layer perceptron-ANN to predict water quality has been demonstrated in recent years (Zhang et al., 2015; García-Alba et al., 2019). However, such studies considered ANN as a component to be integrated within other models and did not present comparisons with other machine learning techniques. Granata et al. (2017) and Haghiabi et al. (2018) used several machine learning methods to forecast water quality and reported a significant overestimation of good water quality but did not explore the reasoning. Moreover, the data sets they used to training the networks were still unbalanced.

It is well known that ANN and other machine learning methods need a good quality training dataset to provide reliable and accurate results. In 2012, Motamarri and Boccelli developed a machine learning model to classify the recreational water quality at a river scale. They found that a neural-based learning vector quantization model outperformed both MLR and ANN, and revealed that all the predictions heavily relied on the validity of the characteristics of the training samples. However, the biggest issue for recreational water quality datasets stay unsloved that they are always unbalanced, with the vast majority of the dataset below guideline thresholds and only a very small amount of data exceeding the guideline thresholds (Thoe et al., 2012; García-Alba et al., 2019). When training data-driven models, an unbalanced dataset increases the possibility of information loss in the minority class and overfitting in the majority class, which cannot be easily addressed by only reducing or duplicating data samples (Batista et al., 2004). Kim et al. (2014) also pointed out to this problem when trying to monitor the water quality using satellite data using three machine learning approaches. The limited training data, especially for bad water conditions, weakened the predictions. To address the unbalanced sample issue, the Synthetic Minority Over-sampling Technique (SMOTE) is often used in other fields (Han et al., 2005; Luengo et al., 2011). SMOTE generates artificial samples by interpolating values between majority and minority classes. However, using this method, the 'synthetic' samples are still more likely to be distributed in the interior of the minority class (Luengo et al., 2011). This limitation can be overcome through the advanced Adaptive Synthetic Sampling algorithm (ADASYN). ADASYN, which can create more samples at the boundary between the two classes and improve training accuracy (He et al., 2008; Gosain and Sardana, 2017). However, in the published ADASYN studies, the data dimension is usually low with no more than three input variables. Therefore, its efficiency on high dimensional data still needs to be assessed. In addition, most studies focus on one site and one model, or at best compare two types of models at one location, usually ANN and linear regression are compared (Thoe et al., 2012; Chan et al., 2013). Thoe et al. (2014) and Danades et al. (2016) addressed the issue by comparing three or more different models to reveal their strengths and weaknesses but only at one location. Other studies focused on multiple locations, but considered only one model (García-Alba et al., 2019). Therefore, a sensitivity analysis using different locations and different algorithms has not yet been performed.

To overcome the aforementioned deficiencies, our study provides a comprehensive comparison of four machine learning algorithms predicting FIB at five beaches in Auckland, New Zealand. The objectives are: (1) to address the problem of unbalanced water quality datasets by applying ADASYN to improve the training of the predictive algorithm; (2) to assess the ability of machine learning models when using a balanced dataset; (3) to compare the different model results at different locations using different machine learning models. By completing these objectives, we try to answer two questions: Can we use artificial water samples to balance the original dataset for machine learning models? and does a balanced dataset improve prediction capability?

2. Study area and data

2.1. Study sites

The five study sites, Milford Beach, Narrow Neck Beach, Judges Bay, Weymouth Beach, and Clarks Beach, are distributed around the coastline of Auckland, New Zealand, and are characterized by distinct geographical conditions (Fig. 1). The first two are open coast beaches, facing the Pacific Ocean and are wave-dominated. The remaining three sites are in sheltered bays and harbours, so are dominated by tidal processes. In 2017, Auckland Council launched a revised model based beach water quality programme ('Safeswim'). The use of models was guided by their ability to meet performance standards published by the United States Geological Survey (USGS – Francy et al., 2013). Based on this guidance, a model with accuracy over 85%, along with sensitivity at 50% and specificity at 80%, should be considered as a qualified model to predict the water quality at a standard level. For a water quality predictive model to be included into the safeswim platform, it must meet the benchmarks in terms of overall accuracy, sensitivity (ability to predict above threshold) and specificity (ability to predict below threshold). Models are in a constant state of review and are refined regularly in response to Auckland Council's ongoing beach sampling.

2.2. Data

Long-term water quality monitoring data is available from 1995 to 2018, with accompanying environmental variables (2019 is also available for Clarks), for the five study sites. Each dataset contains eight variables (Fig. 2): FIB which is used as a threshold to identify the water quality (single sample threshold $= 280$, FIB ≥ 280: exceedance and FIB < 280: compliance) as prescribed in the New Zealand Guidelines, rainfall accumulations for the proceeding twenty four, forty eight, and seventy two hours (one, two, three days), and the total accumulated precipitation, wind direction and speed, and solar hours per day. Fig. 2 shows that Weymouth has the

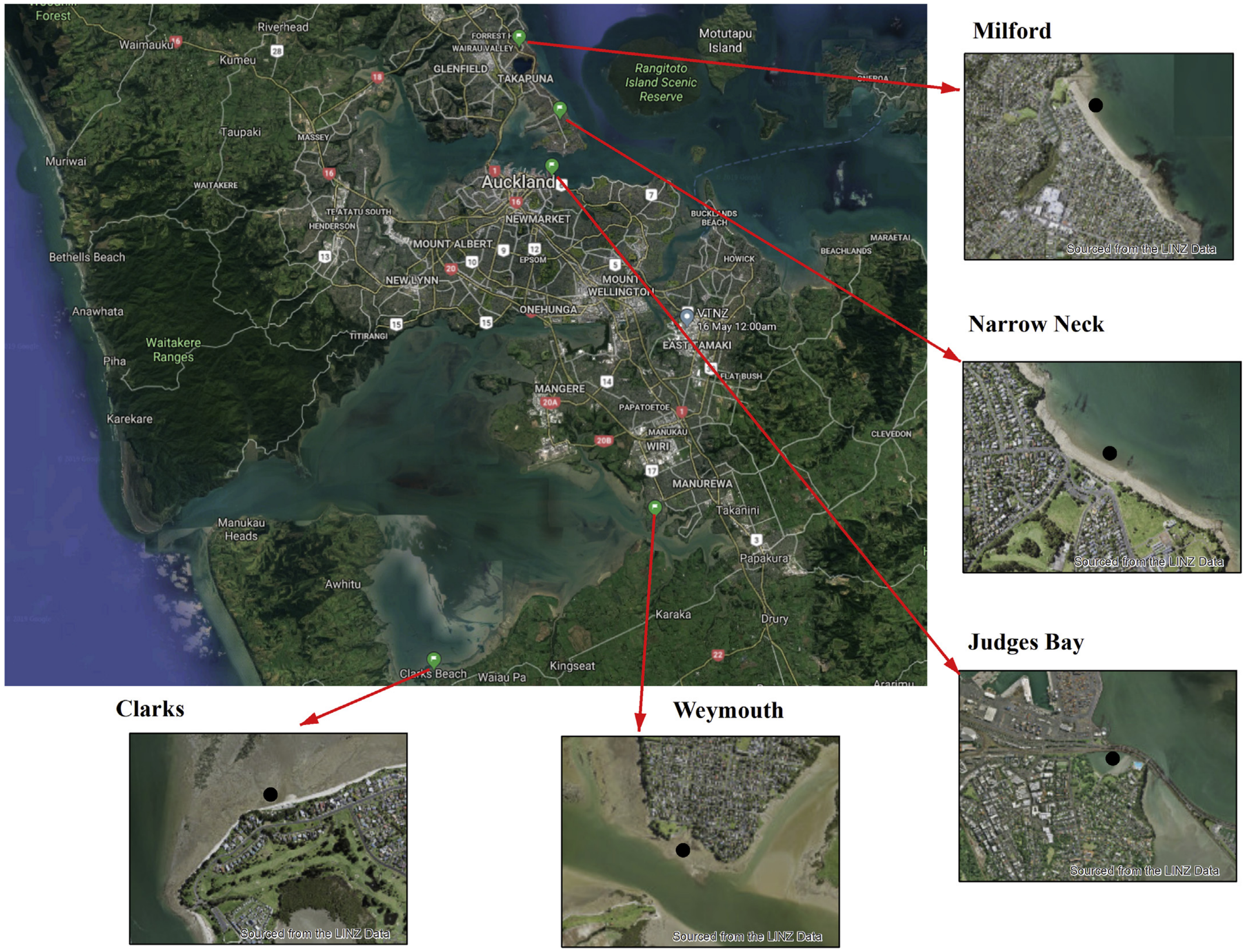


Fig. 1. The five water monitoring sites in Auckland, New Zealand (Black dot in each image indicates the sampling location).

most balanced dataset comparing to other four locations, while Narrow Neck has the most unbalanced dataset, with less than 2.5% samples found to be above the threshold. The wind direction is evenly distributed for Clarks and Weymouth, however, the other three locations experience more southern winds. The average solar hours of Milford is the smallest with about 6.5 h/day and Judges Bay has the longest average sunshine time with almost 7 h/day. The average rainfall accumulation within 72 h in Weymouth is the greatest, almost 8.2 cm, while it is the smallest at Narrow Neck with an average rainfall of 5.6 cm in 72 h (see Fig. 3).

3. Methodology

Unlike regression models that usually predict the absolute FIB value as a continuous function and then reclassify the value based on threshold, four machine learning algorithms are applied to ‘predict’ whether the FIB of a water sample has exceeded the threshold or not through a binary classification process.

3.1. K-nearest neighbors algorithm (K-NN)

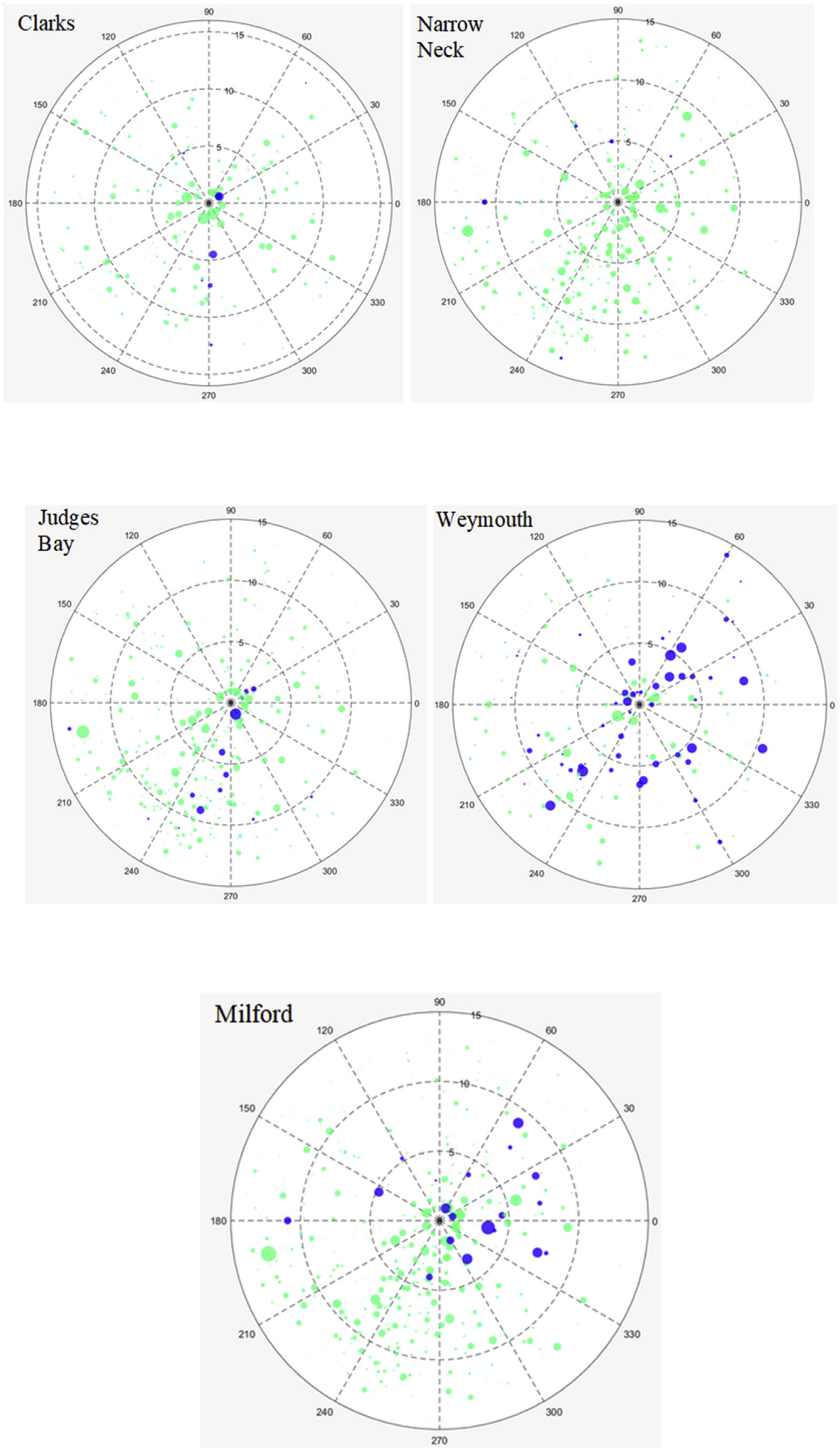
The k-nearest neighbors algorithm is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The neighbors are taken from a set of objects for which the class (for *k*-NN classification) or the object property value (for *k*-NN regression) is known (Fig. 3). This approach is widely used in classification problems (Babbar and Babbar, 2017).

3.2. Boosting decision trees (BDT)

A decision tree is a flowchart-like structure in which each internal node represents a “test” on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules. Boosting is a method that combines many weak learners (trees) into a strong classifier and the Boosting decision trees (Fig. 4) are popular because of their excellent accuracy and fast operation (Prakash et al., 2018; Shoaran et al., 2018).

3.3. Support vector machine (SVM)

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labelled training data, the algorithm outputs an optimal hyperplane


Clarks

Narrow Neck

Judges Bay

Weymouth

Milford

Fig. 2. Datasets at 5 locations in Auckland. Using a polar coordinate system, the direction represents wind direction (0–360°), distance from the origin point represents the daily solar hours (0–15 h), the size of the dot reflects the total precipitation amount accumulating within 72 h (0–100 cm), and the colour indicates whether the water quality exceeded the threshold or not (blue: FIB ≥ 280 and green: FIB < 280). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

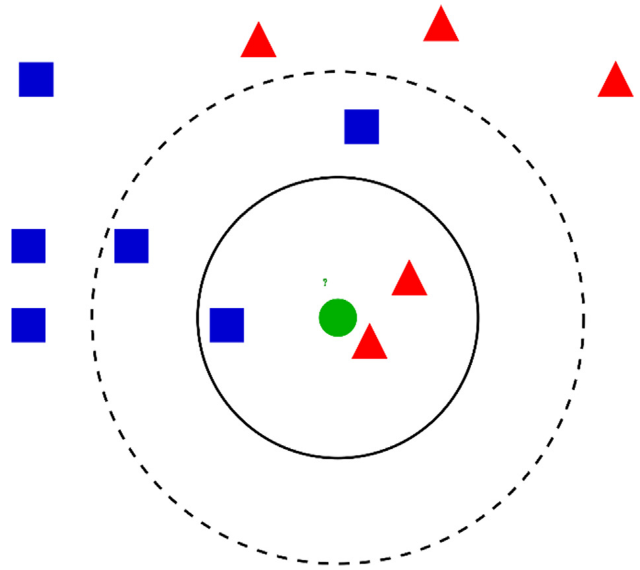


Fig. 3. Example of *k*-NN classification. The test sample (green dot) should be classified either as a blue squares or a red triangles. If *k* = *3* (solid line circle) it is assigned to the red triangle population because there are 2 triangles and only 1 square inside the inner circle. If *k* = *5* (dashed line circle) it is assigned to the blue square population (3 squares vs. 2 triangles inside the outer circle). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

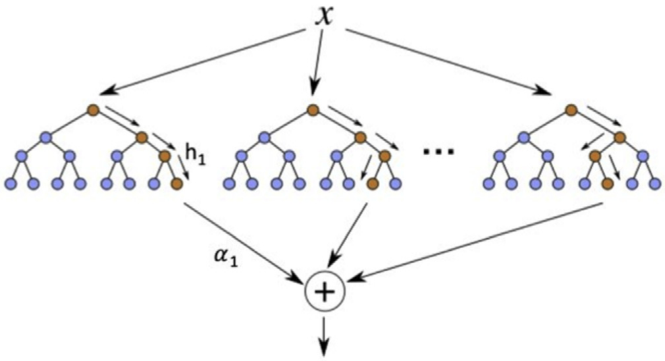


Fig. 4. Boosting decision tree: h is the weak classifier and α represents the extent weight assigned to h.

which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where each class lays on a different side (Fig. 5).

3.4. Artificial neural network (ANN)

ANNs are a widely used modelling technique with self-adapting, self-organizing, and self-learning abilities (Pijanowski et al., 2002, Anctil et al., 2004). ANNs include an input layer (where variables are inputted into the algorithm), one or more hidden layers (where input variables are combined) and an output layer (the prediction). Because of its simplicity, ease of training, and its ability for reasonable associative memory and prediction (Rumelhart et al., 1986), we used a feed-forward, error Back-Propagation Three-Layer Perceptron (BP-TLP) ANN architecture (Fig. 6). The most significant character of a BP-TLP is that during the training stage, the information is transiting forward while the error is back propagating. During the feed-forward transition, the information (values/features) of the input layer is processed by the hidden layer(s) and then transitions back to the output layer. Every node in a layer affects the node state in the following layer(s). If the value of the output layer does not meet the expected outcome, it turns to the back-propagation process. Depending on the expected error, the BP-TLP ANN begins to adjust the weights and threshold values of the network, allowing for the predicted outcome to approach the expected result.

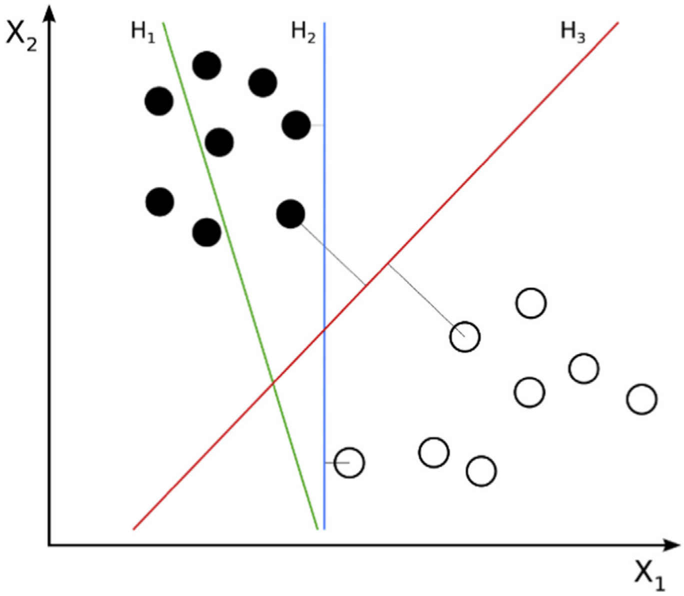


Fig. 5. An example of SVM classification. H1 does not separate the classes. H2 does, but only with a small margin. H3 separates them with the maximum margin.

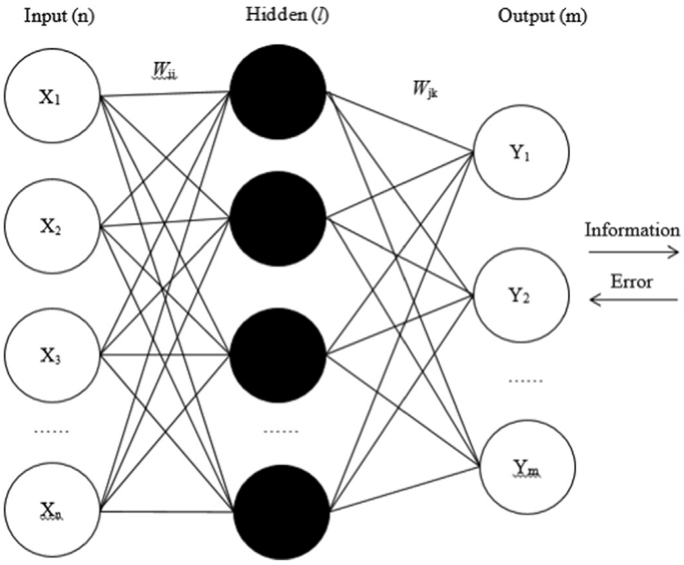


Fig. 6. Architecture of a BP-TLP ANN

3.5. Adaptive synthetic sampling algorithm (ADASYN)

ADASYN is an improved version of the Synthetic Minority Oversampling Technique (SMOTE), which is used to avoid overfitting occurring when exact replicas of minority instances are added to the main dataset (Gosain and Sardana, 2017). The key idea of the ADASYN algorithm is to use the density distribution as a criterion to automatically determine the appropriate number of synthetic samples that need to be generated for each minority data example.

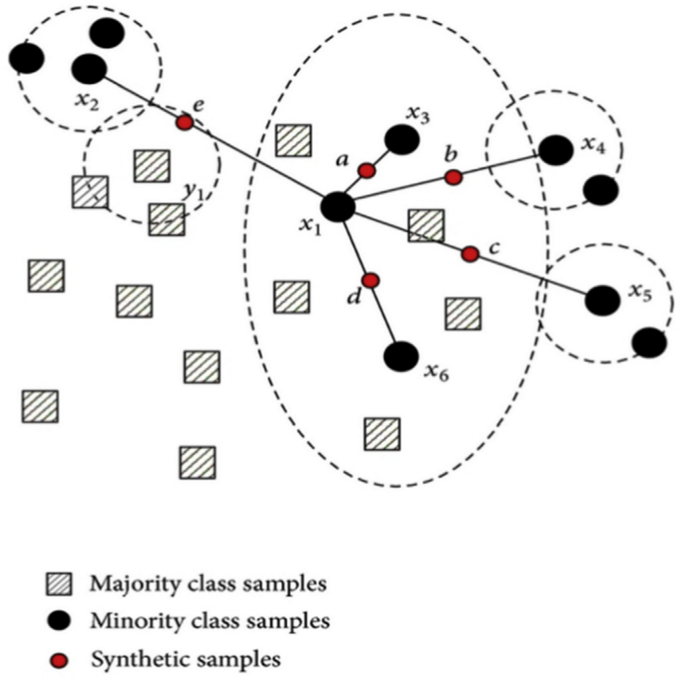


Fig. 7. Generation of Synthetic Instances with the help of SMOTE. a-e are the synthetic samples, x1-x6 are the monitory samples.

The density distribution can be obtained from the knn function based on an n-dimensional vector Euclidean distance between majority and minority samples. A subset of data is taken from the minority class as an example and then new synthetic similar instances are created (Fig. 7). SMOTE stops here but with ADASYN, a random small value is added to the 'faked' samples making the new sample more realistic instead of being linearly correlated to the original sample (more details on the ADASYN application procedures can be found in Appendix I).

3.6. Modelling framework

Fig. 8 illustrates the workflow we used, showing how the data are processed and applied to the models. The sampled data makes the original database, containing two parts as seven training fields and one label field (FIB). Using FIB values, the samples are classified into two classes (FIB above or below 280) to maintain consistency with the New Zealand guidelines. ADASYN is then applied to generate synthetic samples above the threshold to balance the dataset. To form the training dataset, data samples are randomly selected from the samples below the threshold, 280, and combined with the synthesised samples. The new training sample dataset is then fed as input data into the four machine learning algorithms as input data. The unselected remaining samples below the threshold (<280) are coupled with the original (not generated through ADASYN) over-threshold samples (>280) to form the validation and testing dataset and are applied to the validation process of the model results.

The four classifiers, k-nearest neighbourhood (KNN), boosting decision tree (BDT), support vector machine (SVM), and multilayer perceptron (MLP-ANN) were implemented with Matlab. The first three used optimized hyper-parameters with Bayesian optimization process to find the best parameters. The Bayesian optimization is a built-in function of Matlab that can optimize most critical parameters (hyperparameter) automatically. The optimization process can minimize the cross-validation loss by varying the parameter iteratively. For each individual machine learning algorithm, the critical parameters are different. Table 1 describes the critical parameters and the optimization results.

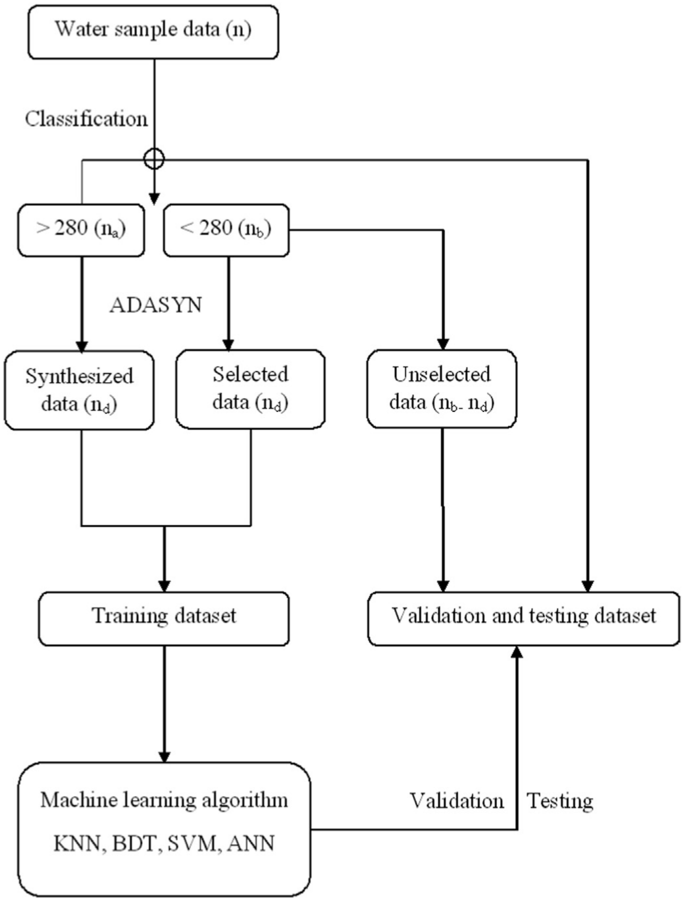


Fig. 8. Flow chart of the modelling framework. In parenthesis is the number of samples. Notice that: $n = n_a + n_b$; $n_d = n_b - n_a$.

While MLP-ANN used different method to decide the parameters. First, the number of hidden nodes significantly affects ANN performance. Too few nodes will cause a significant prediction error, while too many will prolong the training process and lead to overfitting. Based on the rule that the number of hidden neuron should not be less than $2n/3 + 1$ (n = the number of input nodes), we tried the number 6, 7, 8, 9, and the optimal number of hidden neurons, 8, was our final choice in terms of both the model performance and the network simplicity. Then, the Levenberg-Marquardt algorithm function was implemented in Matlab to train the network that updates weights and biases due to its powerful computation with relatively small datasets and the high efficiency for backpropagation. In addition, the transfer functions for hidden layer and output layer are a sigmoid (tansig) and a linear function (purelin), respectively.

Each model was run 100 times for each study site to assess the model capability and stability. A confusion matrix is used to calculate the overall accuracy, sensitivity and specificity, of each model. The running time is also recorded to evaluate the computation efficiency.

Overall, as shown in Fig. 8, for a data D with n samples, n_a and n_b are the number of water samples above and below the FIB threshold, where n_a plus n_b equals n. Notice that n_a is much smaller than n_b and their difference is n_d. In our study, the total demand number of synthesised samples also equals n_d. Then, we randomly selected below-threshold samples from n_b and combine them with

Table 1
Hyperparameters be optimized through Bayesian optimization.

Method Name	Hyperparameter	Description	Result (e.g. Clark)
KNN	Distance metric	Distance searching method of nearest neighbors	mahalanobis
	Neighbour	Number of nearest neighbors to find for classifying each point	8
BDT	Boosting method	Ensemble aggregation method, specially focusing on different boosting functions	GentleBoost
	Number of learning cycle	Total number of learns been trained	43
	Learning rate	The shrinkage range of the learning process	0.0099
	Minimum leaf size	Minimum number of leaf node observations	81
SVM	Box constrain	The maximum penalty imposed on margin-violating observations	0.0089
	Kernel scale	A value used to divide all elements of the predictor matrix	0.001

an equal amount of the synthesised samples to make the new training water sample dataset ($n_{train} = 2n_d$). The original above-threshold samples combined with the remaining below samples (equal to n_b - n_d), are used for model validation. Before applying the modelling framework to the five locations', a simplified under and over sampling method was also applied to reduce the number of below samples and equal the number of below and above samples.

4. Result and discussion

4.1. Original dataset

Table 2 describes the water quality monitoring data samples and Table 3 reveals the training results from all the algorithms using the original dataset. Table 2 shows that all the monitoring sites have a quite unbalanced water samples and only Weymouth has a slightly better balances the number of samples above and below the threshold. According to Table 3, even though the overall accuracy is high, none of the 5 locations can be well modelled with the four ML methods since none of the above threshold samples (exceedances) were successfully predicted. This high model overall accuracy can be problematic because of the unbalance dataset issue. Since all the datasets have very limited occurrences of above threshold samples, algorithms only focus on the below threshold samples. The high overall accuracy of each model is due to the correct prediction of below threshold samples (compliance) while the above samples are not well predicted. Hence, all of these four models can precisely predict the below samples but none of the algorithms can identify the water samples with FIB above the 280 threshold.

4.2. Under/over sampling methods

The under sampling method is first used to balance the data before applying the ADASYN. We reduced the number of below samples through a random selection process until it equalled the number of the above samples. Based on this under sampling dataset, the results (Table 4) indicate that the four aforementioned algorithms can predict the water quality more accurately. KNN and BDT have the same results for all the four locations and predict sample groups from Clarks with a 0.75 overall accuracy. SVM and ANN are also useful to predict the water sample groups from Weymouth and Milford, with an overall accuracy at 0.73 and 0.75. These results make much more sense than using the original data since both the above and below samples can be predicted. However, other models significantly underestimate the below samples at different locations. Taking Milford for example, the SVM underestimated three above samples while overestimated two below samples.

Table 3
Predictions using the raw dataset.

Site	Machine Learning		
	Test Acc.	Above	Below
Clarks	0.93	0/5	68/68
Narrow Neck	0.94	0/8	135/135
Judge Bay	0.94	0/8	132/132
Weymouth	0.87	0/11	73/73
Milford	0.99	0/2	162/162

The underestimation of below samples is due to the loss of potentially useful information with the under sampling method. The training sample may be insufficient as well as it may not be a representative example of the samples below the threshold. Thereby, even the overall accuracy is acceptable, the result is inaccurate and inadequate to be used to predict the water quality.

The over sampling method is the opposite of under sampling. For every location, we duplicated the above samples and then combined them with the below samples to make a new dataset. The results of the over-sampling prediction are shown in Table 5.

In Table 5 we can see that BDT outperformed the other three algorithms in terms of the best overall accuracy, while SVM performed the worst. BDT, KNN, and ANN all display relatively high modelling accuracies at all locations, except for Weymouth. However, results are still deceiving because only replicating the same above samples with the unchanged sample values can increase the likelihood of overfitting. In addition, if there is some error in the samples above threshold, the over sampling method will simply amplify the bias.

4.3. ADASYN approach

Adaptive Synthetic Sampling algorithm (ADASYN) was used to improve data balance. ADASYN synthetically creates new samples from the above samples via linear interpolation. This approach created more samples in the vicinity of the boundary between the two classes than in the interior of the above samples. This approach

Table 2
Water quality samples from the five monitoring sites (above/below threshold).

Threshold = 280	Clarks	Narrow Neck	Judge Bay	Weymouth	Milford
≥280	14	15	23	74	28
<280	231	463	444	207	520
Total	245	478	467	281	548
Ratio (above)	0.0571	0.0314	0.0493	0.2633	0.0511

Table 4
Under-sampling results.

Site	KNN and BDT			SVM			ANN		
	Acc.	Above	Below	Acc.	Above	Below	Acc.	Above	Below
Clarks	**0.75**	3/3	3/5	0.50	3/3	1/5	0.38	3/3	0/5
Narrow Neck	0.56	3/3	2/5	0.33	3/3	0/6	0.33	3/3	0/6
Judge Bay	0.50	6/8	1/6	0.57	8/8	0/6	0.50	6/8	1/6
Weymouth	0.52	11/12	12/32	**0.73**	0/12	32/32	0.43	9/12	10/32
Milford	0.44	6/7	1/9	0.69	4/7	7/9	**0.75**	7/7	5/9

mitigates the problem of overfitting caused by duplicated over-sampling as synthetic new samples are generated artificially so that there is no loss of useful information and less risk of overfitting. Fig. 9 illustrates the dataset obtained with the ADASYN approach.

We first applied the four ML algorithms using the ADASYN dataset and validated and tested with the original dataset for Clarks Bay. The results significantly improve, see Fig. 10 (we show results using MLP ANN as an example, results for other algorithms can be found in Appendix II). In Fig. 10 (1, 1) and (0, 0) represent how many water samples are correctly predicted/classified. The first value indicates the correct prediction for above threshold water samples (sensitivity) and the second value mirrors the accuracy of below sample modelling (specificity)., Overestimation and underestimation are indicated with (0, 1) and (1, 0). The above and below threshold samples are both precisely predicted by KNN using the ADASYN dataset for training and the original dataset for validation. The overall accuracy is 88.43%. The accuracy of specificity (below samples) and sensitivity (above samples) is 89.25% and 76.47%, respectively, both are higher than the under sample methods and other models. With this balanced dataset, the underestimation and overestimation are also acceptable since only 10.75% of the above threshold samples were missed and 23.53% of below samples were false alarmed to be above. Compared to other studies, the specificity accuracy does not increase significantly but the sensitivity prediction accuracy improved to over 70%.

The difference between over-sampling and ADASYN is the data sample itself. For over-sampling, we simply duplicate the existing minority class samples (above 280) until the number equals to the majority class (below 280). For example, assume there are 200 samples below the threshold value, 280, and only 20 samples with a value greater than 280. The simplified over-sampling will only copy these 20 above samples for 10 times until reaching the same number as the below samples and these duplicated samples will feed into the training and validation process. Therefore, the final model validation result could be really high simply because of using the same data samples in both training and validation process. In addition, this approach magnifies overfitting.

However, with the ADASYN method, the dataset will be balanced with brand new artificial above-threshold data samples by producing new values rather than simply replicating the existing ones. There is no conflict between training and validation dataset. The synthetic samples will only be used to train the model and the true data will only be applied to the model validation.

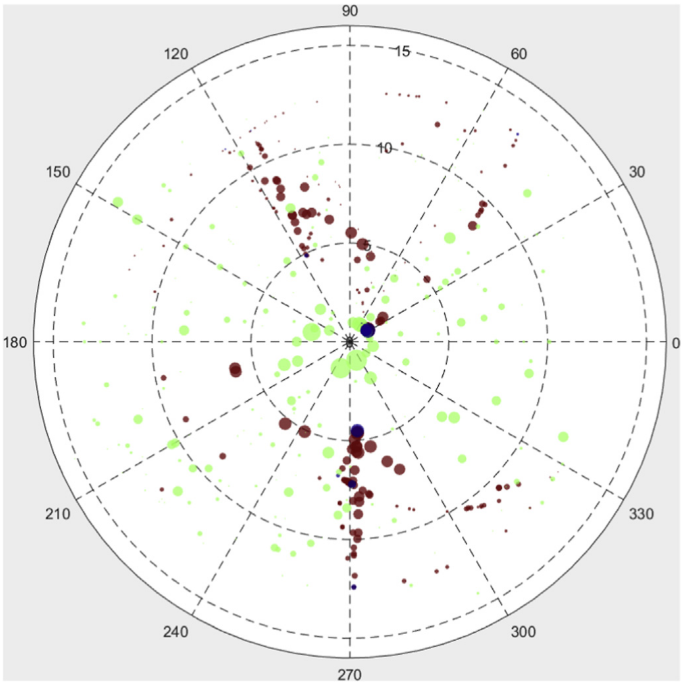
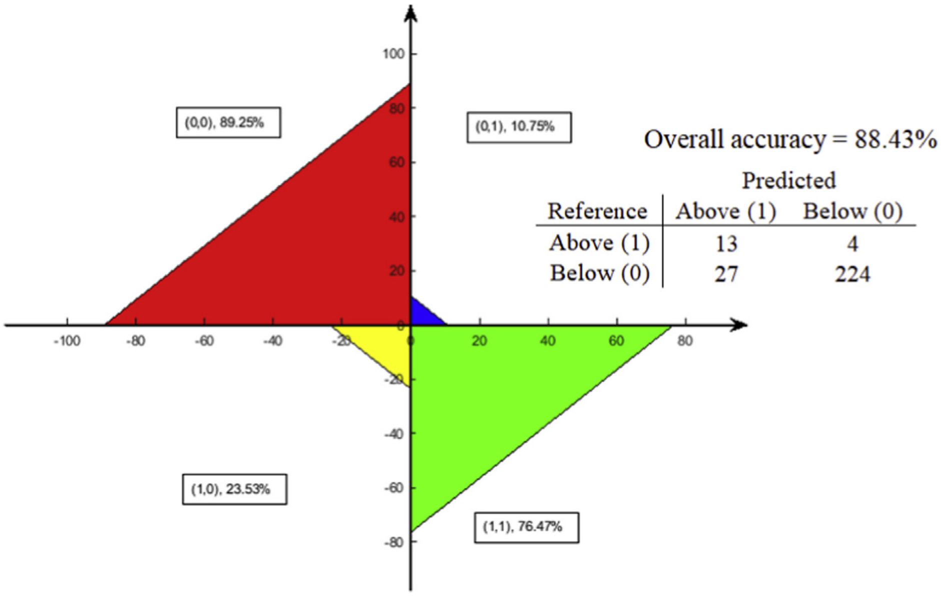


Fig. 9. Balanced data for Clarks Beach, green dot: below sample, blue dot: above sample, brown dot: ADASYN sample. Direction, distance from the center, and size of the dots represent the wind direction (0–360°), daily solar hours (0–15 h), and the total precipitation amount accumulating within 72 h (0–100 cm). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Based on above explanations, even if the accuracy value of over-sampling is higher than ADASYN, we cannot say that the over-sampling model outperforms the ADASYN due to the data duplication issue. In contrast, the results from ADASYN are more reliable because the dataset is balanced and separates the training and validation datasets by creating artificial data.

Table 5
Over-sampling results.

Site	KNN			BDT			SVM			ANN		
	Acc.	Above	Below	Acc.	Above	Below	Acc.	Above	Below	Acc.	Above	Below
Clarks	0.98	72/72	66/69	**0.99**	64/64	76/77	0.69	46/68	51/73	0.94	59/59	43/49
Narrow Neck	0.98	147/147	126/132	**0.99**	142/142	136/137	0.68	105/142	85/137	0.97	102/102	105/112
Judge Bay	0.96	127/127	128/138	**0.99**	132/132	131/133	0.67	64/131	114/134	0.91	80/80	104/123
Weymouth	0.83	65/68	42/61	**0.88**	62/68	51/61	0.64	47/73	36/56	0.73	31/47	41/52
Milford	0.98	154/154	151/157	**0.99**	163/163	145/148	0.73	93/151	135/160	0.97	109/109	123/130



	Predicted	
Reference	Above (1)	Below (0)
Above (1)	13	4
Below (0)	27	224

Fig. 10. Visualization of the MLP-ANN result using ADASYN, Clarks Beach. Results for other algorithms are presented in Appendix II.

4.4. *Comparison four model results among the five locations*

Fig. 11 (at the end of this article) displays the validation results between models and original data for all the locations using four ML algorithms with ADASYN dataset. Boxplot diagrams, used to reveal the model accuracy (we performed 100 runs with each algorithm), show the highest accuracy (top line), the mean accuracy (middle line), the lowest accuracy (bottom line), and the accuracy range (box size). The red crosses indicate any eventual outlier in the model predictions.

KNN precisely predicted the above and below samples for all the beaches with an average accuracy around 80%. Except for Weymouth, at least one KNN can predict the above and below samples correctly, nearly reaching a 100% accuracy. However, the box size of KNN is always larger than the other three models (the lowest accuracy for each site is always associated to KNN), indicating that a large variation existed for the 100 runs of KNN. We conclude that the results are not as robust and reliable as the other models. SVM is the algorithm showing the worst performance with very low mean model accuracy values, never higher than 0.7. This model should not be considered as an appropriate prediction model because of its inferior performance at all the locations (some of the SVM based

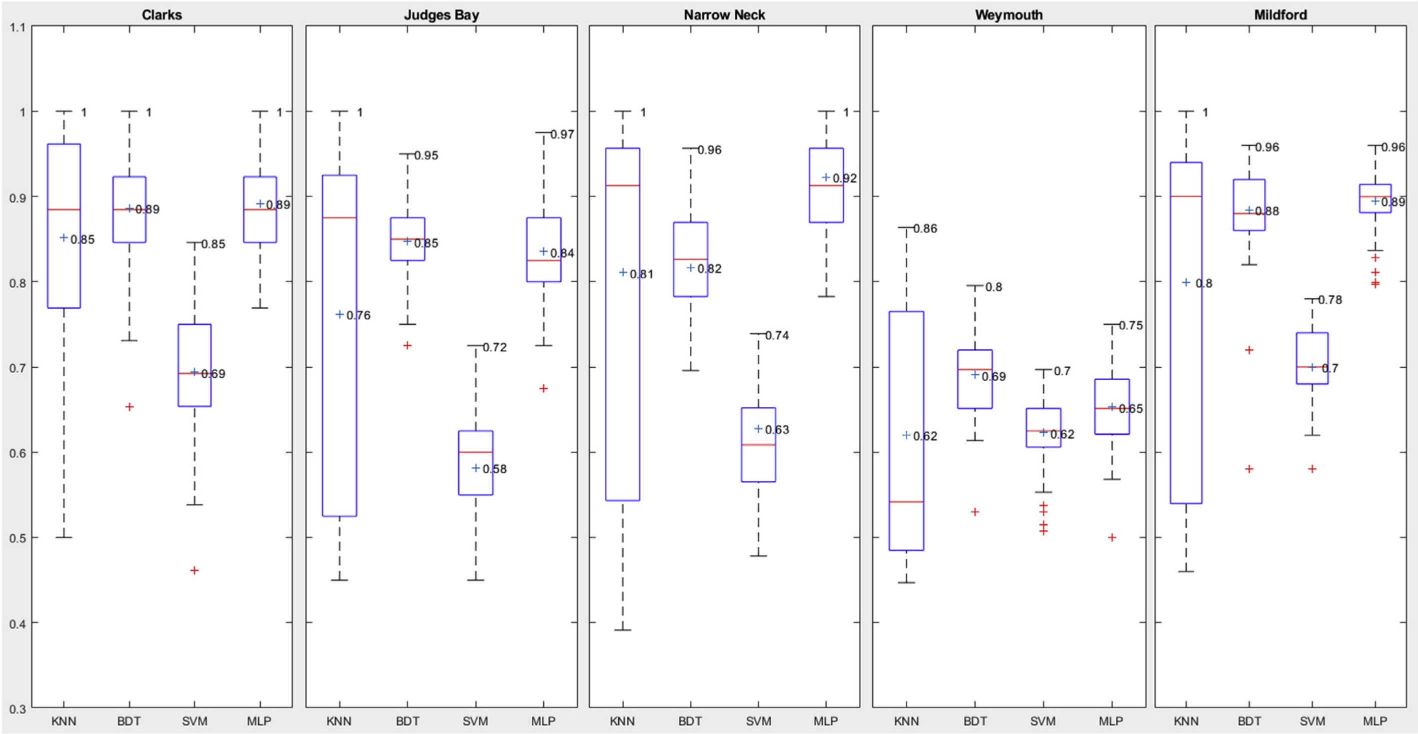


Fig. 11. Boxplots showing the modelling accuracy at the five monitoring sites.

predictions are below 0.5). BDT and MLP-ANN predictions are less variable than KNN and SVM (relatively smaller box sizes in Fig. 11). The results from BDT and MLP-ANN are very similar for Clarks, Weymouth, and Milford. BDT outperformed ANN slightly at Judges Bay (smaller box size, higher mean, 0.85, and minimum accuracy values, 0.70), however, the ANN provides a better result at Narrow Neck with a much higher mean accuracy value (0.92) than BDT (0.82).

At different location, model performance is also different. Both BDT and MLP can be used to predict the above and below water samples for Clarks Beach (high average accuracy, 0.89, and stability, small box size). For Judges Bay, BDT should be considered as the most appropriate model because of the second highest mean accuracy value (0.85) and least variation among all the models. For Narrow Neck, MLP has the highest mean accuracy (0.92), nearly 10% higher than the BDT (0.83), and the results are stable. Weymouth has the worst modelled results compared to other locations. The best prediction in Weymouth is obtained using KNN, while the BDT has the highest mean accuracy with relatively stable results so that overall they could both be taken into consideration when predicting at this location. Similar to Clarks, both BDT and MLP can provide solid results for Milford with almost the same highest and mean accuracy (0.96/0.89). However, MLP is more stable than BDT with a smaller box size and a higher minimum accuracy value so that it is the first choice for predicting above and below water samples at this location.

Our results demonstrate that there is no unique model that fits all sites and the reason can be attribute to the different conditions and settings of each location (see also Shaw et al., 2017). Different models should always be tested and the results should be compared to look for the most appropriate solution. Model choice also depends on which effect of the results is considered. For example, if we are only concerned with model accuracy, KNN can be used as it provides the best results and can predict the above and below samples with nearly a 100% accuracy. If we care more about the model stability, KNN then should be eliminated and BDT and MLP-ANN should be considered because of their relatively low variation between lowest and highest accuracy. Even SVM could be considered as it is the most stable model for Weymouth regardless the accuracy.

ADASYN created a balanced samples dataset by generating ‘artificial’ above samples through a linear interpolation at the boundary between below and above samples. With the half-faked data set, the accuracy of model predictions significantly improved. However, this approach may raise some other issues which need to be further addressed. For example, the artificial samples could either be invading or expanding too much to the below class, which will cause overfitting problem (Luengo et al., 2011).

5. Conclusion

To help resolve the issue of unbalanced datasets in water quality models, which result in loss of information on the above threshold samples (exceedance) and overfitting for the below threshold sample (compliance), we used an ADASYN sample balancing method to generate artificial data of above threshold samples. Together with machine learning techniques, this sample balancing resulted in more accurate predictions of water quality, compared with the original unbalanced datasets. The balanced dataset was applied to four machine learning algorithms and resulted in much higher accuracy in terms of both sensitivity (over 75%) and specificity (over 90%) compared to past studies (Thoe et al., 2014; Zhang et al., 2015). K-NN, BDT, SVM, and MLP-ANN were applied to five beach sites in Auckland and the results were compared to determine which model is the most appropriate for each individual location. Judging in terms of model accuracy, robustness and stability, BDT and MLP-ANN outperformed KNN and SVM at all locations.

Models perform differently at different locations and many reasons could explain this result (e.g., environmental conditions, data quality, and model suitability). This study provided a potential solution to the data balance issue and compared various models to test the predictions at five sites. Future studies could focus on the role of individual variables or on reducing uncertainties in the synthetic generation of data. In addition, the tidal information, which is very important for predicting FIB, is not included in this study due to lack of detailed data. Tidal data should be added into future models and it is likely it would further improve model performance.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We want to show special thanks to the Auckland Council for suppling of water quality data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.watres.2020.115788.

References

Abyaneh, H.Z., 2014. Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. J. Environ. Health Sci. Eng. 12 (1), 40.

Anctil, F., Perrin, C., Andréassian, V., 2004. Impact of the length of observed records on the performance of ANN and of conceptual parsimonious rainfall-runoff forecasting models. Environ. Model. Software 19 (4), 357—368.

Babbar, R., Babbar, S., 2017. Predicting river water quality index using data mining techniques. Environ. Earth Sci. 76 (14), 504.

Bae, H.K., Olson, B.H., Hsu, K.L., Sorooshian, S., 2010. Classification and regression tree (CART) analysis for indicator bacterial concentration prediction for a California coastal area. Water Sci. Technol. 61 (2), 545e553.

Batista, G.E., Prati, R.C., Monard, M.C., 2004. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter 6 (1), 20—29.

Bedri, Z., Corkery, A., O'Sullivan, J.J., Deering, L.A., Demeter, K., Meijer, W.G., O'Hare, G., Masterson, B., 2016. Evaluating a microbial water quality prediction model for beach management under the revised EU Bathing Water Directive. J. Environ. Manag. 167, 49e58.

Castelletti, A., Yajima, H., Giuliani, M., Soncini-Sessa, R., Weber, E., 2014. Planning the optimal operation of a multioutlet water reservoir with water quality and quantity targets. J. Water Resour. Plan. Manag. 140 (4), 496—510. https://doi.org/10.1061/(ASCE)WR.1943—5452.0000348.

Chan, S.N., Thoe, W., Lee, J.H.W., 2013. Real-time forecasting of Hong Kong beach water quality by 3D deterministic model. Water Res. 47 (4), 1631—1647.

Chandramouli, V., et al., 2007. Backfilling missing microbial concentrations in a riverine database using artificial neural networks. Water Res. 41 (1), 217—227.

Chen, W.B., Liu, W.C., 2015. Water quality modeling in reservoirs using multivariate linear regression and two neural network models. Adv. Artif. Neural Syst. 6, 2015.

de Brauwere, A., Ouattara, N.K., Servais, P., 2014. Modeling fecal indicator bacteria concentrations in natural surface waters: a review. Crit. Rev. Environ. Sci. Technol. 44 (21), 2380e2453.

Danades, A., Pratama, D., Anggraini, D., Anggriani, D., 2016. October. Comparison of accuracy level K-nearest neighbor algorithm and support vector machine algorithm in classification water quality status. In: 2016 6th International Conference on System Engineering and Technology (ICSET). IEEE, pp. 137—141.

Francy, D.S., Brady, A.M.G., Carvin, R.B., Corsi, S.R., Fuller, L.M., Harrison, J.H., Hayhurst, B.A., Lant, J., Nevers, M.B., Terrio, P.J., Zimmerman, T.M., 2013. Developing and Implementing Predictive Models for Estimating Recreational Water Quality at Great Lakes Beaches. U.S. Geological Survey Scientific

Investigations, p. 68. https://doi.org/10.3133/sir20135166/. Report 2013-5166.

García-Alba, J., et al., 2019. Artificial neural networks as emulators of process-based models to analyse bathing water quality in estuaries. Water Res. 150, 283–295.

Granata, F., Papirio, S., Esposito, G., Gargano, R., De Marinis, G., 2017. Machine learning algorithms for the forecasting of wastewater quality indicators. Water 9 (2), 105.

Gosain, A., Sardana, S., 2017. September. Handling class imbalance problem using oversampling techniques: a review. In: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, pp. 79–85.

Haghiabi, A.H., Nasrolahi, A.H., Parsaie, A., 2018. Water quality prediction using machine learning methods. Water Qual. Res. J. 53 (1), 3–13.

He, H., Bai, Y., Garcia, E.A., Li, S., 2008. June. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE, pp. 1322–1328.

He, L.M.L., He, Z.L., 2008. Water quality prediction of marine recreational beaches receiving watershed baseflow and stormwater runoff in southern California. Water Res. 42 (10–11), 2563–2573. USA.

Han, H., Wang, W.Y., Mao, B.H., 2005. August. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: International Conference on Intelligent Computing. Springer, Berlin, Heidelberg, pp. 878–887.

Kazemi Yazdi, S., Scholz, M., 2010. Assessing stormwater detention systems treating road runoff with an artificial neural network predicting fecal indicator organisms. Water Air Soil Pollut. 206 (1e4), 35e47.

Kim, Y.H., Im, J., Ha, H.K., Choi, J.K., Ha, S., 2014. Machine learning approaches to coastal water quality monitoring using GOCI satellite data. GIScience Remote Sens. 51 (2), 158–174.

King, R.S., Richardson, C.J., 2003. Integrating bioassessment and ecological risk assessment: an approach to developing numerical water-quality criteria. Environ. Manag. 31 (6), 795–809.

Li, Y., Khan, M.Y.A., Jiang, Y., Tian, F., Liao, W., Fu, S., He, C., 2019. CART and PSO+KNN algorithms to estimate the impact of water level change on water quality in Poyang Lake, China. Arab. J. Geosci. 12 (9), 287.

Luengo, J., Fernández, A., García, S., Herrera, F., 2011. Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary undersampling. Soft Computing 15 (10), 1909–1936.

Marshall, M.M., Naumovitz, D., Ortega, Y., Sterling, C.R., 1997. Waterborne protozoan pathogens. Clin. Microbiol. Rev. 10 (1), 67–85.

Motamarri, S., Boccelli, D.L., 2012. Development of a neural-based forecasting tool to classify recreational water quality using fecal indicator organisms. Water Res. 46 (14), 4508e4520.

Pijanowski, B.C., et al., 2002. Using neural networks and GIS to forecast land use changes: a land transformation model. Comput. Environ. Urban Syst. 26 (6), 553–575. https://doi.org/10.1016/S0198-9715(01)00015-1.

Prakash, R., Tharun, V.P., Devi, S.R., 2018. April. A comparative study of various classification techniques to determine water quality. In: 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). IEEE, pp. 1501–1506.

Qin, X., Gao, F., Chen, G., 2012. Wastewater quality monitoring system using sensor fusion and machine learning techniques. Water Res. 46 (4), 1133–1144.

Rumelhart, D., Hinton, G., Williams, R., 1986. Learning internal representations by error propagation. In: Rumelhart, D.E., McClelland, J.L. (Eds.), Parallel Distributed Processing: Explorations in the Microstructures of Cognition, vol. 1. MIT Press, Cambridge, pp. 318–362.

Shoaran, M., Haghi, B.A., Taghavi, M., Farivar, M., Emami-Neyestanak, A., 2018. Energy-efficient classification for resource-constrained biomedical applications. IEEE J. Emerg. Sel. Top. Circuits Syst. 8 (4), 693–707.

Shaw, A.R., et al., 2017. Hydropower optimization using artificial neural network surrogate models of a high-fidelity hydrodynamics and water quality model. Water Resour. Res. 53 (11), 9444–9461.

Stidson, R.T., Gray, C.A., McPhail, C.D., 2012. Development and use of modelling techniques for real-time bathing water quality predictions. Water Environ. J. 26 (1), 7e18.

Thoe, W., et al., 2014. Predicting water quality at Santa Monica Beach: evaluation of five different models for public notification of unsafe swimming conditions. Water Res. 67, 105–117.

Thoe, W., Lee, J.H.W., 2013. Daily forecasting of Hong Kong beach water quality by multiple linear regression (MLR) models. ASCE J. Environ. Eng., 04013007 https://doi.org/10.1061/(ASCE)EE.1943-7870.0000800.

Thoe, W., Wong, S.H.C., Choi, K.W., Lee, J.H.W., 2012. Daily prediction of marine beach water quality in Hong Kong. J. Hydro-Environ. Res. 6 (3), 164–180.

Tufail, M., et al., 2008. Artificial intelligence-based inductive models for prediction and classification of fecal coliform in surface waters. J. Environ. Eng. 134 (9), 789–799.

Wade, T.J., Calderon, R.L., Brenner, K.P., Sams, E., Beach, M., Haugland, R., Wymer, L., Dufour, A.P., 2008. High sensitivity of children to swimming-associated gastrointestinal illness: results using a rapid assay of recreational water quality. Epidemiology 375–383.

Wang, X., Zhang, J., Babovic, V., 2016. Improving real-time forecasting of water quality indicators with combination of process-based models and data assimilation technique. Ecol. Indicat. 66, 428e439.

Zhang, Z., et al., 2015. Modeling fecal coliform bacteria levels at Gulf coast beaches. Water Qual. Expo. Health 7 (3), 255–263.