Water Research 171 (2020) 115454

Contents lists available at ScienceDirect

Water Research

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

Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data

Kangyang Chen [a, 1], Hexia Chen [b, 1], Chuanlong Zhou [b], Yichao Huang [b], Xiangyang Qi [a], Ruqin Shen [b, c], Fengrui Liu [d], Min Zuo [e], Xinyi Zou [a], Jinfeng Wang [c], Yan Zhang [c], Da Chen [b], Xingguo Chen [a, f, **], Yongfeng Deng [b, c, *], Hongqiang Ren [c]

[a] Jiangsu Key Laboratory of Big Data Security & Intelligent Processing, Nanjing University of Posts and Telecommunications, Nanjing, China
[b] School of Environment, Guangzhou Key Laboratory of Environmental Exposure and Health, Guangdong Key Laboratory of Environmental Pollution and Health, Jinan University, Guangzhou, Guangdong, 510632, China
[c] State Key Laboratory of Pollution Control and Resource Reuse, School of the Environment, Nanjing University, Nanjing, Jiangsu, 210023, China
[d] College of Literature, Science, and the Arts, University of Michigan, Ann Arbor, MI, 48109, USA
[e] National Engineering Laboratory for Agri-product Quality Traceability, Beijing Technology and Business University, Beijing, Beijing, 100048, China
[f] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, Jiangsu, 210023, China

ARTICLE INFO

Article history:
Received 9 August 2019
Received in revised form
24 December 2019
Accepted 30 December 2019
Available online 31 December 2019

Keywords:
Water quality prediction
Machine learning models
Ensemble methods
Deep cascade forest
The key water parameters

ABSTRACT

The water quality prediction performance of machine learning models may be not only dependent on the models, but also dependent on the parameters in data set chosen for training the learning models. Moreover, the key water parameters should also be identified by the learning models, in order to further reduce prediction costs and improve prediction efficiency. Here we endeavored for the first time to compare the water quality prediction performance of 10 learning models (7 traditional and 3 ensemble models) using big data (33,612 observations) from the major rivers and lakes in China from 2012 to 2018, based on the precision, recall, F1-score, weighted F1-score, and explore the potential key water parameters for future model prediction. Our results showed that the bigger data could improve the performance of learning models in prediction of water quality. Compared to other 7 models, decision tree (DT), random forest (RF) and deep cascade forest (DCF) trained by data sets of pH, DO, CODMn, and NH_3−N had significantly better performance in prediction of all 6 Levels of water quality recommended by Chinese government. Moreover, two key water parameter sets (DO, CODMn, and NH3−N; CODMn, and NH3−N) were identified and validated by DT, RF and DCF to be high specificities for perdition water quality. Therefore, DT, RF and DCF with selected key water parameters could be prioritized for future water quality monitoring and providing timely water quality warning.

1. Introduction

There have been ever increasing applications of artificial intelligence (AI) technologies in many fields, benefiting various works of our life. For examples, face recognition technologies based on deep learning models are able to track the criminals precisely and timely (LeCun et al., 2015). Other sought-after applications including artificial intelligence driving technologies in some complex driving environment (Aeberhard et al., 2015), and global mega-size medical data analyses using deep learning (Esteva et al., 2019; Gurovich et al., 2019). In the environmental monitoring fields, especially hydrological monitoring, AI technologies have also started to be successfully applied using various learning algorithm models. For instance, the real-time radar-derived rainfall forecasting were proposed by some machine learning models (Yu et al., 2017). Moreover, An intelligent water regimen monitoring system were

* Corresponding author. School of Environment, Guangzhou Key Laboratory of Environmental Exposure and Health, Guangdong Key Laboratory of Environmental Pollution and Health, Jinan University, Guangzhou, Guangdong, 510632, China.
** Corresponding author.
E-mail addresses: chenxg@njupt.edu.cn (X. Chen), yongfengdeng@jnu.edu.cn (Y. Deng).
[1] Kangyang Chen, Hexia Chen contributed equally to this work.

https://doi.org/10.1016/j.watres.2019.115454
0043-1354/

List of abbreviations

COD_{Mn}	Chemical oxygen demand
CRF	Completely-random tree forest
CRT	Completely-random tree
DCF	Deep cascade forest
DO	Dissolved oxygen
DT	Decision tree
KNN	K-nearest neighbors
LDA	Liner discriminant analysis
LR	Logistic regression
NB	Naive bayes
NH_3—N	Ammonia-nitrogen
pH	Hydrogen ion concentration
RF	Random forest
SVM	Support vector machines

recently tried to established to solve measuring accuracy problem, and increase working efficiency (Fan et al., 2018).

Additionally, application of AI dealing with large-scale environmental medium monitoring data have also gradually gained interests. With regards to hydrological monitoring, a great number of environmental monitoring data have traditionally been collected worldwide by individual environmental departments and agencies (Dunbabin and Marques, 2012). Consolidating these huge environmental data sets would provide valuable resources for training machine learning models in order to identify patterns and markers for understanding for instance water quality. Moreover, although the environmental quality (water or air) could be classified and reported real-time by monitoring reference parameters according to the guidelines of environmental classification, they are unable to provide early warning of future environmental pollution. Importantly, machine learning models could be trained starting with a few key environmental quality parameters and the final environmental outcomes using the current archived data, then to predict the future outcomes after inputting new parameters. This suggested the possibility of using some environmental parameters to quickly predict future environmental consequences (Ghahramani, 2015). It is of great significance to be able to provide accurate environmental early warning with minimum possible environmental parameters yet maximum prediction precision of the learning models. Despite the great potential of machine learning in the area of environmental medium quality monitoring, the relevant research and applications of machine learning models are still lacking.

Surface water is a kind of non-renewable resource and of great significance in our human daily life. However, with the development of the economy, surface water quality continues to be compromised and deteriorate posing threats to human health especially in some developing countries (Vörösmarty et al., 2010; Zhang et al., 2010). As the largest developing country in the world, In the last decades, many parts of China were facing have induced severe water stress as well as the health risks water contamination, after ever-growing demands and misuse of surface water resources over the last decades (Sun et al., 2016). Therefore, it is particularly crucial to monitor and predict surface water quality precisely and timely. Indeed, a few machine learning methods were created and implemented to monitor and predict water quality, after reviewing related previous studies in the past decade (Table S1). However, most of them used relatively small training and validation datasets. More importantly, these traditional models demonstrated relatively low prediction precision and unbalanced prediction abilities for different Levels of water qualities. For example, the highest prediction precision of water quality for succeeding weeks in the above Avila et al. study was 92.00% for Green (Acceptable), 0.00% for Amber (Alert), and 95.00% for Red (Action) predicted from bayesian network model (Avila et al., 2018). Indeed, the prediction precision for Green (Acceptable) and Red (Action) action may be satisfied for the governors, however, the prediction precision for Amber (alert) is not acceptable. For Environmental Protection Agencies, they often should take some measures as soon as possible when facing the alert of water quality deterioration. Undoubtedly, the water quality prediction precision of these older models would be limited them for application in the future, and the precision of those learning models may be limited by few training data in these previous studies. Nevertheless, the relation between surface water quality performance and the amount of training data was still largely unknown.

Recently, an emerging class of machine learning models named ensemble methods were proposed, which consisted of multiple learning algorithms (Zhang and Ma, 2012). Importantly, they often had better predictive performance than any one of the constituent learning algorithms alone (Qiu et al., 2017; Singh et al., 2013). Random forest (RF) as one of the most famous representatives have been proposed and applied in various areas (Belgiu and Drăguţ, 2016; Boulesteix et al., 2012). Notably, gcForest, a decision tree ensemble method, proposed by Zhou et al., in 2017 also attracted much attention (Zhou and Feng, 2018). It was a new deep learning model called deep cascade forest (DCF) by cascading several random forests using the idea of ensemble learning. Compared to the earlier ensemble and deep learning models, DCF could adaptively determine the number of cascade levels and set the model complexity automatically enabling excellent performance even on small-scale data (Zhou and Feng, 2018), so users could control training costs according to the available computational resources. Furthermore, the performance of DCF is quite robust in hyper-parameter settings when changing different data from different domains, however, the other deep learning models such as deep neural networks was deeply required tuning network structure and hyper-parameters (Zhou and Feng, 2018). Surprisingly, there was limited information on the prediction of surface water quality using DCF. Moreover, there was little information about the comparative analysis of the prediction performance from different ensemble learning models based on the same big data.

Notably, the water quality prediction performance of machine learning models may be not only dependent on the models per se and the amount of data set, but also dependent on the water parameters in data set chosen for training the learning models. It is vital to identify some key water parameters for training learning models without significantly reducing predictive performance, because they could remarkably increase prediction efficiency and decrease the prediction cost. However, few studies have evaluated the surface water quality prediction performance of learning models to identify key water parameters from big training data sets (Singh and Kaur, 2017).

The aims of this study were to: **i**) based on the precision, recall, F1-score and weighted F1-socre, investigate whether the big data could improve the water prediction performance of 7 traditional machine learning models (The 7 models with high frequency in previous studies, Table S1) [logistic regression (LR), linear discriminant analysis (LDA), support vector machine (SVM), decision tree (DT), completely-random tree (CRT), naive bayes (NB), k-nearest neighbors (KNN)] and 3 emerging ensemble learning models [random forest (RF), completely-random tree forest (CTF), and deep cascade forest (DCF)] using the weekly data of national large rivers and lakes (33,612 observations) collected from China National Environmental Monitoring Centre from 2012 to 2018; **ii**)

comprehensively compare the water prediction performance of these 10 learning models by using the same big training and validation sets, and identify the best model for future surface water quality monitoring and providing timely water quality warning; **iii**) identify the key water parameters using the machine learning model with significantly better prediction performance in above experiments for further increasing prediction efficiency and decreasing the prediction cost.

2. Materials and methods

2.1. Study area and water quality data

In this study, the raw data (33,614 observations) were collected weekly data form 124 automatic water quality monitoring stations in 10 national large rivers and lakes from China National Environmental Monitoring Centre from 2012 to 2018 (http://www.mee.gov.cn/hjzl/shj/dbszdczb//). The 10 national large rivers and lakes were distribution throughout the country, including Songhua River, Liaohe River, Haihe River, Huaihe River, Yellow River, Yangtze River, Pearl River, Taihu Lake, Chaohu Lake, and Dianchi Lake, and their geographical locations were presented in Fig. S1.

Four selected water quality parameters, including chemical oxygen demand (COD_{Mn}), dissolved oxygen (DO), ammonia-nitrogen (NH_3—N), and hydrogen ion concentration (pH), and their overall water quality classification results were adopted directly from China National Environmental Monitoring Centre and were used for constructing water quality prediction models. Based on the National Environmental quality standards for surface water of China (GB3838-2002), water quality was classified into five Levels (I, II, III, IV, and V). Additionally, the water quality worse than V was also considered in this study. Hence, here the Level of water quality was classified and predicted using six Levels from good to bad: I, II, III, IV, V, and worse than V (WV), respectively.

Here 33,612 water quality data sets were obtained after data clean (remove the data sets with missing values). Additionally, in order to keep the all water parameters with the same degree of influence on final outcomes, here we performed *Z*-score standardization on the water parameters (Barboza et al., 2017) (Fig. 1).

2.2. Seven traditional machine learning models

In order to evaluate whether big training data could improve the water quality prediction performance by machine learning models, the seven traditional machine learning models (DT, NB, LR, LDA, CRT, KNN, and SVM) with higher frequency in previous studies (Table S1) were selected in this study (Fig. 1). And the detailed information about these seven older models were presented in the supporting information.

2.3. Three ensemble learning models

For the comprehensive comparing the water quality prediction performance between traditional and ensemble learning models and identifying the potential models for further water quality monitoring, RF, CRF, and DCF as the representatives of ensemble learning models were selected in this study (Fig. 1).

2.3.1. RF

RF is an ensemble learning model based on DT as base models (Breiman, 2001). In the RF, each tree uses a sample obtained by bootstrap (Gislason et al., 2006). Each node has d features of the base DT. It randomly selects $\sqrt{d}$ features from the feature set which has d features as a new subset, and then selects the optimal feature by *GiNi* value from the subset for split (Strobl et al., 2008). By voting all the decision trees' output, we get the final output (Fig. S2).

Where C_k^i is the predictive class of $tree_i$, after using majority voting, class C_k is the final predictive class of $\vec{x}$. The process of majority voting is as follows: If there are K classes to be predicted, the prediction of x by $tree_i$ in RF is represented as an K-dimensional vector $\vec{c_i} = (h_i^1, h_i^2, \ldots, h_i^K)$, where $h_i^k \in [0, 1]$, indicates the predictive probability of class C_k of $tree_i$. Briefly, the voting method used by RF with T decision trees is defined as:

$$y = \begin{cases} C_k, & \text{if } \sum_{i=1}^{T} h_i^k > 0.5 \sum_{k=1}^{K} \sum_{i=1}^{T} h_i^k; \\ \text{reject}, & \text{otherwise.} \end{cases} \quad (1)$$

The prediction will take effect only if more than half of the trees predict the same result, and here 10 DT were selected in the forest.

2.3.2. CRF

Similar to RF, the CRF was an ensemble learning model based on CRT as base models (Breiman, 2001), and the 10 CRT were selected to construct the CRF in this study.

2.3.3. DCF

As a highly ensemble method, deep cascade forest can be constructed by many random forests (Zhou and Feng, 2018). In each layer of deep cascade forest, we can also add other traditional machine learning models and ensemble models to improve the ensemble performance.

Each level of cascade receives feature information processed by its preceding level, and outputs its processing result to the next level. In addition, to reduce the risk of over-fitting, all the base models in each layer use k-fold cross validation. When a layer is trained, we achieved the predictive accuracy on a validation set. The model will continue to construct the next layer until the accuracy can be no longer improved. Consequently, the number of cascade levels can be automatically determined by the termination of the training process. The adaptive adjustment of model complexity enables deep cascade forest to be applied to training data sets of different scales.

Here we increased the diversity at full stretch. Firstly, we selected two different types of forests (2 RF and 2 CRF) to encourage the diversity in each layer of DCF (Fig. S3) so that the input feature perturbation and data sample perturbation could be increased. Next, we voted to combine all the individual output together and generated the final output that increases the data sample perturbation. Based on different amounts of input features, we also increased the input feature perturbation. Since, the input of next layer came from the output of previous layer and the raw features in our model, we definitely increased the output representation perturbation. In all, DCF is a decision tree ensemble model with high diversity and low generalization error.

2.4. Model operation and evaluation

2.4.1. Model operation

The all 10 model operation was carried out in python 3.6 with different packages (Fig. 1). For DT, NB, LR, LDA, CRT, KNN, SVM, RF and CRF were built by package of Scikit-learn v.019 (Pedregosa et al., 2011) using big training data (33612 sets of 4 water parameters). DCF was built by package of gcForest v.1.1.1 (Zhou and Feng, 2018). It has been well demonstrated that the two packages could be used to successfully build the all 10 learning models with high-level performance (Pedregosa et al., 2011; Zhou and Feng, 2018). After building models, the main structures, functions and hyper-parameters were obtained and presented in Table 1.

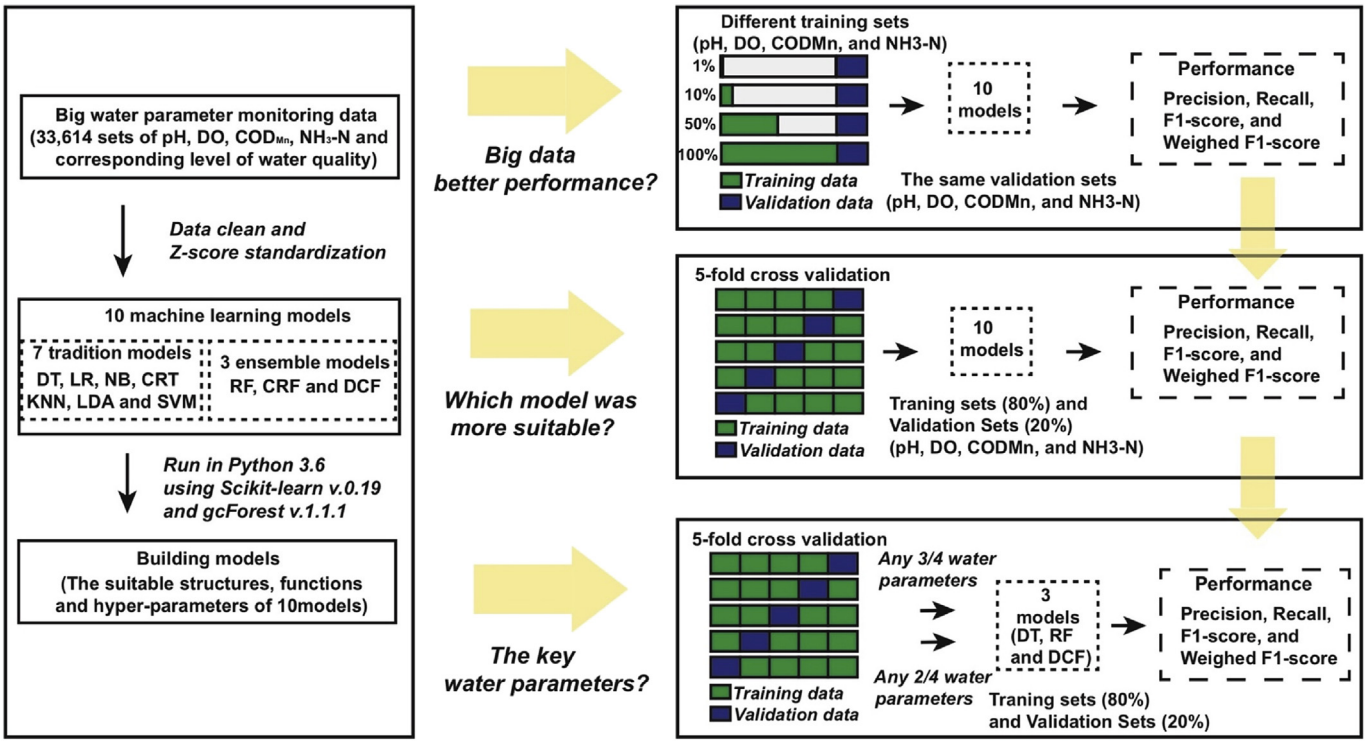

Fig. 1. The sketch map of this study.

2.4.2. Model evaluation

The precision and recall were selected for evaluating the prediction performances of learning models (Fig. 1). Moreover, F1-score and their weighted average value (weighed F1-score) were also selected for better assessing this multi-classification problem (Fig. 1). These 4 metrics were defined as:

The *precision* was defined as:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

The *recall* was defined as:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

The F1-score was defined as:

$$F1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (4)$$

The weighted F1-score was defined as:

Table 1
The main functions and hyper-parameters of 10 machine learning models.

Learning Models	Model main functions and hyper-parameters													
	Penalty	C[a]	Solver	Kernel	Gamma[d]	Criterion	Min Sample Split[f]	Min Sample Leaf[g]	Max Features	Neighbors	Estimators	Bootstrap	Max Layers	Early Stopping Rounds
LR	L2	1	Liblinear	–	–	–	–	–	–	–	–	–	–	–
LDA	–	–	SVD[b]	–	–	–	–	–	–	–	–	–	–	–
SVM	–	1	–	RBF[c]	0.25	–	–	–	–	–	–	–	–	–
DT	–	–	–	–	–	Gini[e]	2	1	N	–	–	–	–	–
CRT	–	–	–	–	–	Gini[e]	2	1	Sqrt(N)	–	–	–	–	–
NB	–	–	–	Gaussian	–	–	–	–	–	–	–	–	–	–
KNN	–	–	–	–	–	–	–	–	–	5	–	–	–	–
RF	–	–	–	–	–	Gini[e]	2	1	Sqrt(N)	–	10	True	–	–
CRF	–	–	–	–	–	–	2	1	Sqrt(N)	–	–	False	–	–
DCF	L2	1	Liblinear	–	–	Gini[e]	2	1	N/Sqrt(N)[h]	–	10	True	10	3

[a] Penalty parameter of the error term.
[b] Singular value decomposition.
[c] Radial Basis Function Kernel.
[d] The coefficient of Radial Basis Function Kernel.
[e] Gini impurity.
[f] The minimum number of samples required to split an internal node.
[g] The minimum number of samples required to be at a leaf node.
[h] DCF contained 4 RF, 3 DT and 1 LR, and N for DT, Sqrt(N) for RF, respectively.

$$Weighted\ F1-score=\frac{1}{K}\sum_{i=1}^{K}\beta_i\cdot F1_i \quad (5)$$

Where *TP* is true positive (positive class is predicted as positive class); *FP* is false positive (negative class is predicted as positive class); *FN* is false positive (positive class is predicted as negative class); *K* is the number of classes, and β_{ii} is the proportion of the number of samples in class *i* to the total number of samples. $F1_i$ is the F1 score of X in class *i*.

Additionally, a stratified *k*-fold ($k = 5$) cross validation was used to estimate the prediction performance of models (Fig. 1). The advantage of this method is that all observations are used for both training and validation.

3. Results

3.1. The training data used in this study

After reviewing the all clean data (33,612 observations), the Level I, I, II, III, IV, V, WV were 7.25% (2438 observations), 39.49% (13,272 observations), 26.17% (8797 observations), 16.28% (5472 observations), 4.42% (1487 observations), and 6.38% (2146 observations) (Fig. 2A). Notably, the water quality including Level IV, V, WV was 33.47% (9105 observations), which means the water quality prediction for Chinese larger rivers was necessary. For the four fundamental water parameters, pH (7.68 ± 0.54), DO (8.10 ± 2.60 mg/L), COD_{Mn} (4.26 ± 3.35 mg/L), and NH_3—N (0.56 ± 1.42 mg/L) ranged from 5.79 to 10.19, 0.02 to 123, 0 to 110.80, and 0 to 30.1, respectively (Fig. 2B). Each water parameter was standardized using their respective Z-score prior to inputting training models in order to keep all the 4 parameters possessing the same degree of influence on final water quality prediction (Fig. 2C).

3.2. Water quality prediction performances after increasing training data

In order to evaluate whether better water quality perdition performances of each learning models could be observed after increasing training data, here the data sets were firstly divided into two parts including training (80%, 26890 observations) and validation sets (20%, 6722 observations) (Fig. 1). The training set was further divided into subsets of different proportions [(1%, 269 observations), (10%, 2689 observations), (50%, 13455 observations), and (100%, 26890 observations)] for training these 10 models subsequently (Fig. 3A), while the validation data set remained the same. The main functions and hyper-parameters of these 10 models were presented in Table 1. In the training process, the traditional learning models including KNN and SVM presented better performance after increasing training data, whereas water quality prediction performance of 3 ensemble learning models did not show any improvement with increasing training data in the training process, based on precision, recall and F1-socre (Fig. S4 and Table S2). Importantly, here the all learning models, expect for LDA, showed improved prediction performance in the validation process after inputting more training data (Fig. S5 and Table S3). Notably, the highest improvement for prediction performance of all 9 models were through inputting 10% of the training data compared with the performance of these 9 models trained by 1% training data, based on the weighted F1-socre (Fig. 3B), and the improving prediction performance of these 9 models were form 1.87% (LR) to 22.76% (KNN). Interestingly, the higher average improvement of water quality prediction were identified in 6 traditional models (11.62%) compared to that in 3 ensemble models (7.30%) (Fig. 3B). Although the prediction performance of these 9 models were still increasing, there were limited improvement of these 9 models after further increasing training data from 10% to 100% (Fig. 3B).

3.3. The ensemble learning models presented better water quality prediction performance compared to traditional learning models

In order to identify the potential suitable models for future water quality monitoring, we further compared the performance of 7 traditional learning models with 3 ensemble learning models (Fig. 1). We used four water fundamental parameters based on the maximum training data (80%, 26890 observations), because most learning models had better performance after increasing training data. The main functions and hyper-parameters of these 10 models were kept the same with the above experiments (Table 1). The performance of 10 learning models for all 6 Levels of water qualities in both training and validation processes were evaluated by precision, recall, F1-score and weighted F1-score (Fig. 1). Additionally, a stratified *5*-fold cross validation was also used to estimate the stability of prediction performance of models (Figs. 1 and 4A).

3.3.1. The training process

In the training process, amongst 7 traditional models, the DT and CRT had the highest precision (1.00 ± 0.00 for DT, 1.00 ± 0.00 for CRT, Fig. 4B and Table S4) recall (1.00 ± 0.00 for DT, 1.00 ± 0.00 for CRT, Fig. 4B and Table S4), and F1-score (1.00 ± 0.00 for DT, 1.00 ± 0.00 for CRT, Fig. 4B and Table S4) in predicting water quality at all 6 Levels. However, the LR, LDA, SVM, NB and KNN presented unbalanced abilities for predicting 6 Levels of water qualities (Fig. 4B and Table S4). Specially, the LR and LDA failed to predict the

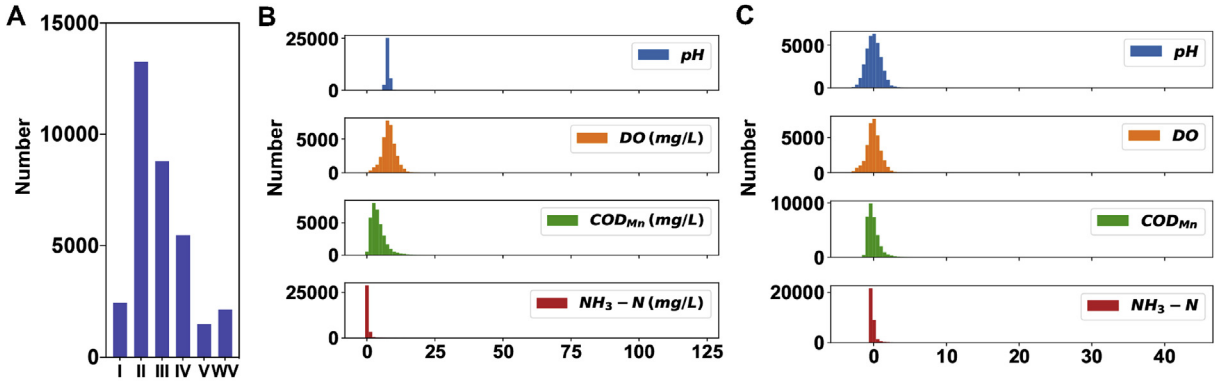

Fig. 2. The data used in this study. (A) Water quality; (B) 4 water parameters; (C) The data after *Z*-score standardization.

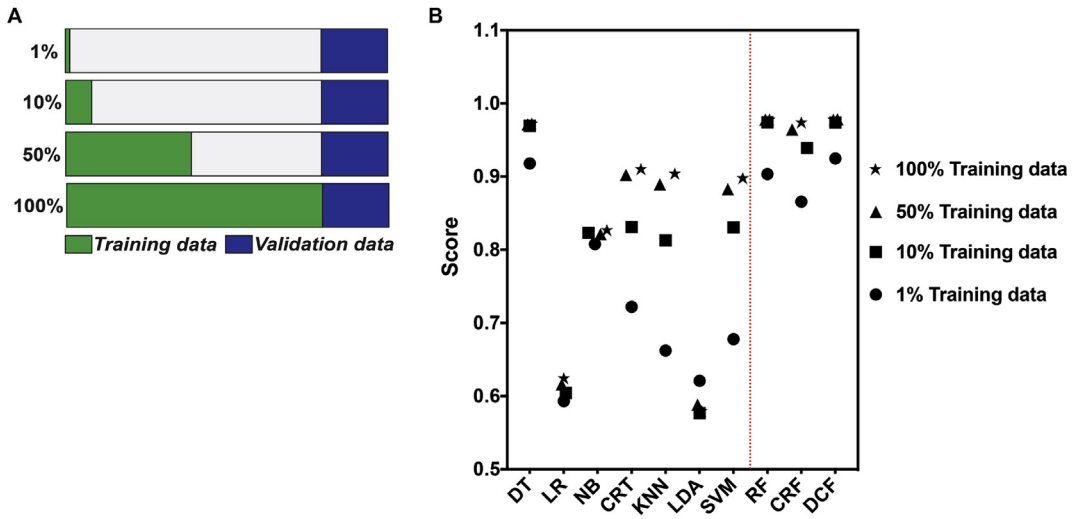

Fig. 3. The surface water quality prediction performance of 10 learning models in the validation process after increasing training data. (A) The sketch map of amount of training data; (B) Weighted F1-socre.

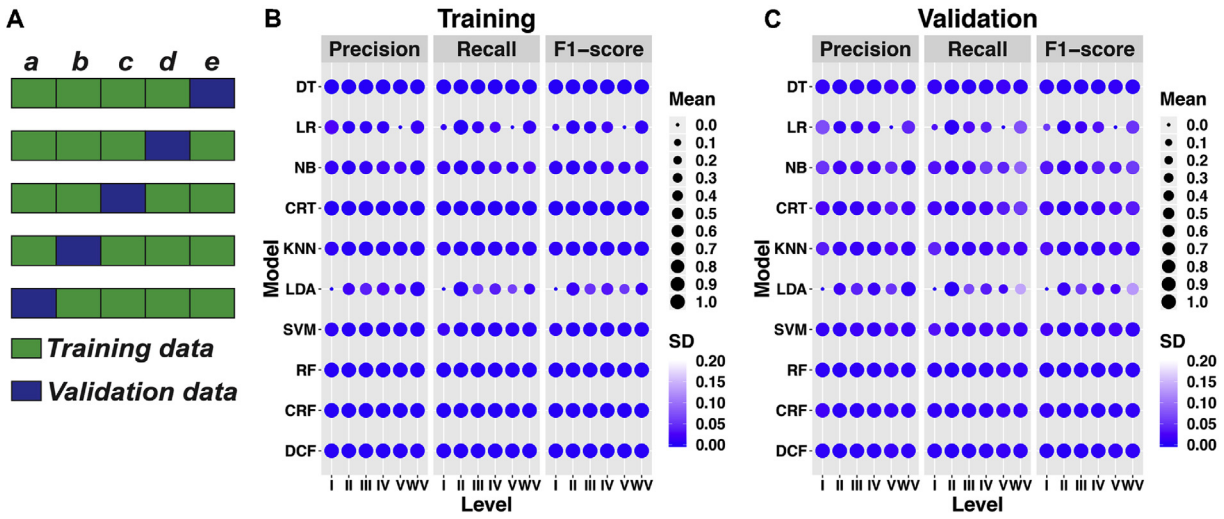

Fig. 4. The surface water quality prediction performance of 10 learning models after training big data. (A) The sketch map of 5-fold cross validation; (B) The prediction performance of models in training process; (C) The prediction performance of models in validation process.

Level V and Level I water, respectively. The SVM presented significantly worse performance for predicting the Level I water than other 5 Levels water ($P < 0.001$), whereas NB presented significantly worse performance for predicting the Level V water than other 5 Levels water ($P < 0.001$). Interestingly, KNN also showed significant reduction in performance for predicting the Level I and Level V water than other 3 Levels water ($P < 0.001$).

Next, the 3 ensemble learning models including RF, CRF, and DCF were also trained by the same data sets. Overall, all 3 ensemble learning models in this study showed satisfactory performance for predicting all 6 Levels of water qualities in training processes (Fig. 4B and Table S4). Similar to DT and CRT, CRF showed better performance than RF and DCF did in the processes of training, which were with the highest precision (1.00 ± 0.00) recall (1.00 ± 0.00), and F1-score (1.00 ± 0.00) for predicting all 6 Levels of water qualities. Notably, all 3 ensemble learning models showed balanced and comprehensive abilities for predicting 6 Levels of water qualities (Fig. 4B and Table S4), which were better than the above LR, LDA, SVM, NB and KNN. Importantly, the RF, CRTF, and DCF showed significantly better performance for predicting Level I and Level V water than LR, LDA, SVM, NB and KNN ($P > 0.001$). Additionally, no more significant difference was observed between RF and DCF predicted all 6 Levels of water qualities during the training process.

3.3.2. The validating process

After training, the 20% additional data were selected for validation the 10 learning models (Figs. 1 and 4A). Compared to other 6 tradition learning models, DT still showed remarkably better performance for predicting all 6 Levels of water qualities water in validation process ($P < 0.001$, Fig. 4C and Table S5). The LR and LDA presented unacceptable performance for predicting Level I and Level V water in validating process (Fig. 4C and Table S5). Notably, the performance of CRT presented different trend between the processes of validating and training. Although it had the best performance during training process, CRT presented significant worse performance for predicting all 6 Levels of water qualities water in validating process ($P < 0.001$, Fig. 4C and Table S5). Additionally, the

performance of other 3 learning models including SVM, NB and KNN in validating process was similar with their did in training process, which also showed significantly worse performance for predicting the Level I and Level V water than other 3 Levels water ($P < 0.01$, Fig. 4C and Table S5).

Compared to these 7 traditional learning models, the all 3 ensemble learning models in our study shown excellent performance for predicting all 6 levels of water qualities in validating processes (Fig. 4C and Table S5). Among the 3 ensemble learning models, RF and DCF presented better performance than CRF for predicting all 6 Levels of water qualities, especially in predicting the Level V and Level WV water ($P < 0.001$). Notably, RF, CRF and DCF presented significantly better predicting performance of all 6 Levels of water qualities compared to all 7 traditional learning models except that DT selected in this study. More importantly, RF and DCF showed satisfactorily performance for predicting the Level V and Level WV water, which had highest precision [0.98 ± 0.01 (Level V), 0.99 ± 0.00 (Level WV) for RF; 0.99 ± 0.01 (Level V), 0.99 ± 0.01 (Level WV) for DCF], recall [0.99 ± 0.01 (Level V), 0.99 ± 0.01 (Level WV) for RF; 0.99 ± 0.01 (Level V), 0.99 ± 0.01 (Level WV) for DCF,] and F1-score [0.98 ± 0.01 (Level V), 0.99 ± 0.01 (Level WV) for RF; 0.99 ± 0.01 (Level V), 0.99 ± 0.01 (Level WV) for DCF]. Finally, based on the weighted F1-score, the DCF and CRF shown the significant better water quality prediction performance compared to these 7 traditional learning models except DT (Fig. 5).

3.4. The water quality prediction performance of DT, RF and DCF based on different water quality parameters sets

For identify the key water parameters, here the DT, RF and DCF would be selected again for training and validating the prediction these water qualities based on any 3 and 2 of the 4 water quality parameters (Fig. 1), due to DT, RF, and DCF has best performance for prediction all 6 Levels of water qualities based on 4 water quality parameters.

3.4.1. Any 3 of 4 water quality parameters

Generally, the performance of DT, RF and DCF for predicting all 6 Levels of water qualities in both training (Fig. S6 and Table S6) and validating processes (Fig. 6 and Table S7) based on any 3 water parameters were worse than that DCF trained by 4 water quality parameters. Nevertheless, the DT, RF and DCF trained by DO, CODMn, and NH3—N still presented acceptable performance for prediction water qualities, which shown highest precision, recall

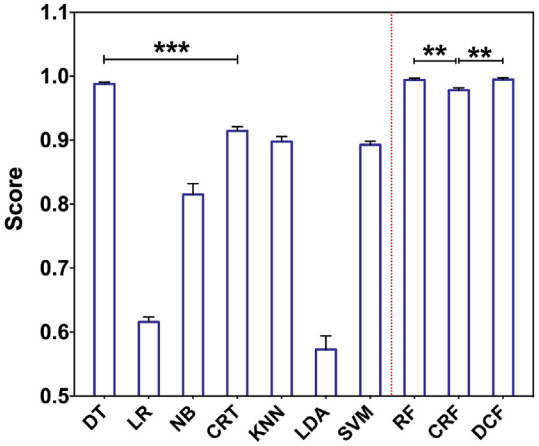

Fig. 5. The weighted F1-score of 10 learning models in the validation process after training big data.

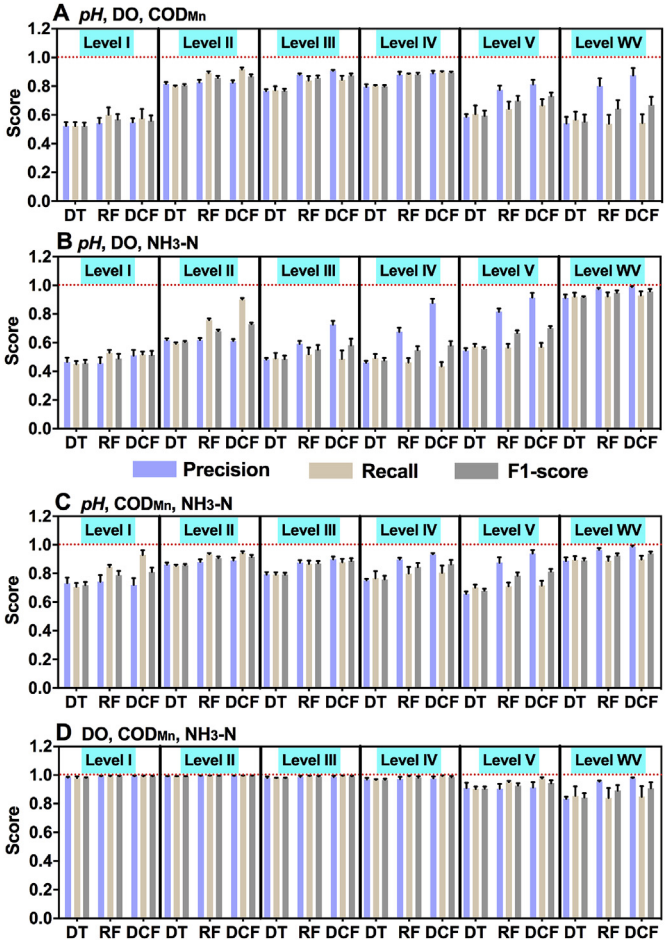

Fig. 6. The surface water quality prediction performance of 10 learning models in the validation process using 3 water parameters after training big data. (A) pH, DO, COD_{Mn}; (B) pH, DO, NH_3—N; (C) pH, COD_{Mn}, NH_3—N; (D) DO, COD_{Mn}, NH_3—N.

and F1-score for predicting the water qualities form Level I to V in validating processes (Fig. 6 and Table S7). Although the performance for predicting the Level WV water from these 3 models trained by DO, CODMn, and NH3—N was not significant higher than the 3 models trained by the other 3 parameter sets (pH, CODMn, NH_3—N; pH, DO, NH_3—N; pH, DO, CODMn), the lowest precision, recall, F1-score from DT were 0.83 ± 0.02, 0.85 ± 0.07, 0.84 ± 0.03 (Fig. 6D and Table S7), respectively, which were also accepted in water quality prediction. Additionally, the weighted F1-score of DT, RF and DCF trained by DO, CODMn, and NH_3—N were 0.97 ± 0.00, 0.98 ± 0.01, 0.98 ± 0.01 and significant higher ($P < 0.001$) that of these 3 models trained by the other 3 water parameter sets (Fig. 7).

3.4.2. Any 2 of 4 water quality parameters

DT, RF and DCF were then selected for training (Fig. S7 and Table S8) and validation of the prediction these water qualities using any 2 of the 4 water quality parameters. Similarly, the performance of DT, RF and DCF based on any 2 water parameters in validation process were worse than that these 3 models trained by 3 or 4 water quality parameters (Fig. 8 and Table S9). The performance of DT, RF and DCF trained by CODMn and NH_3—N was significantly better than that these 3 models trained by other 5 training sets (pH, DO; pH, CODMn; pH, NH_3—N; DO, CODMn; and DO, NH_3—N), and their weighted F1-score was 0.86 ± 0.01, 0.86 ± 0.01, 0.88 ± 0.01(Fig. 9).

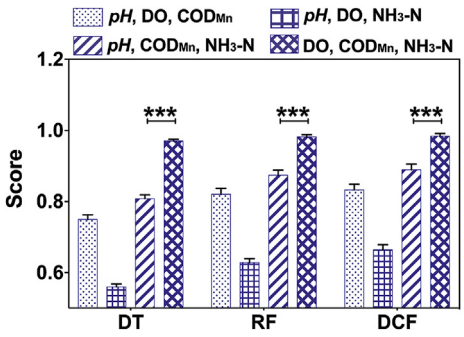

Fig. 7. The weighted F1-score of 10 learning models in the validation process using 3 water parameters after training big data.

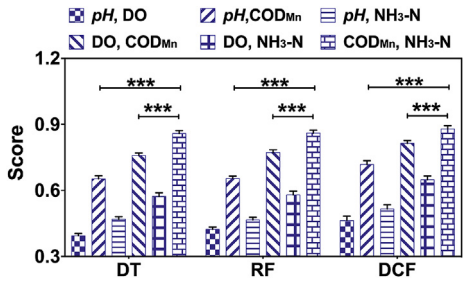

Fig. 9. The weighted F1-score of 10 learning models in the validation process using 2 water parameters after training big data.

4. Discussion

It has been well established that surface water quality depends on multiple water conditions such as COD_{Mn}, DO, NH_3—N, pH, temperature, total phosphorus, metal ion and other toxic chemicals (Simeonov et al., 2003). Meanwhile, chlorophyll-α and *E. coli* were also chosen as direct biological indicators for evaluating the water quality (Avila et al., 2018; Park et al., 2015), because they can characterize the ecological state and illness risks of the tested water. However, they were simplistic and couldn't provide the direct factors deteriorated water quality. Therefore, a set of fundamental water parameters including the above COD_{Mn}, DO, NH_3—N, pH, water temperature, total phosphorus, metal ion and serval toxic organic chemicals have been recommended for assessing water quality by several National Government Departments (Alabaster and Lloyd, 2013). For example, 24 fundamental water parameters were selected for classifying the water quality, according to the guideline of National Environmental quality standards of surface water of China (GB3838-2002) (MEEPRC, 2002). Based on the results of all 24 basic water parameters, it could classify water quality accurately and quickly by some lookup procedures. However, it is costly and time consuming to measure all 24 water parameters on all kinds of surface water, due to the various experimental requirements. Therefore, it is more practical to measure key parameters indicative of water quality rather than completely following the guideline of GB3838-2002 to understand water quality. Hence, it is of great significance to predict water quality with fewer but more indicative fundamental water parameters.

Increasing number of studies have tried to predict the surface water quality by machine learning models using a few water parameters and of the reported water quality (Di et al., 2019; Tan et al., 2012), however, the relatively poor prediction precision of these learning models still limits their further application. It is most likely that these learning models used relatively small amount of training data, because recent studies have shown that better performance could be achieved with more data used during the training process in each of the learning models (Chen et al., 2016; Zhu et al., 2012). In our study, improving performance of water quality prediction using 9 different learning models were observed after increasing training

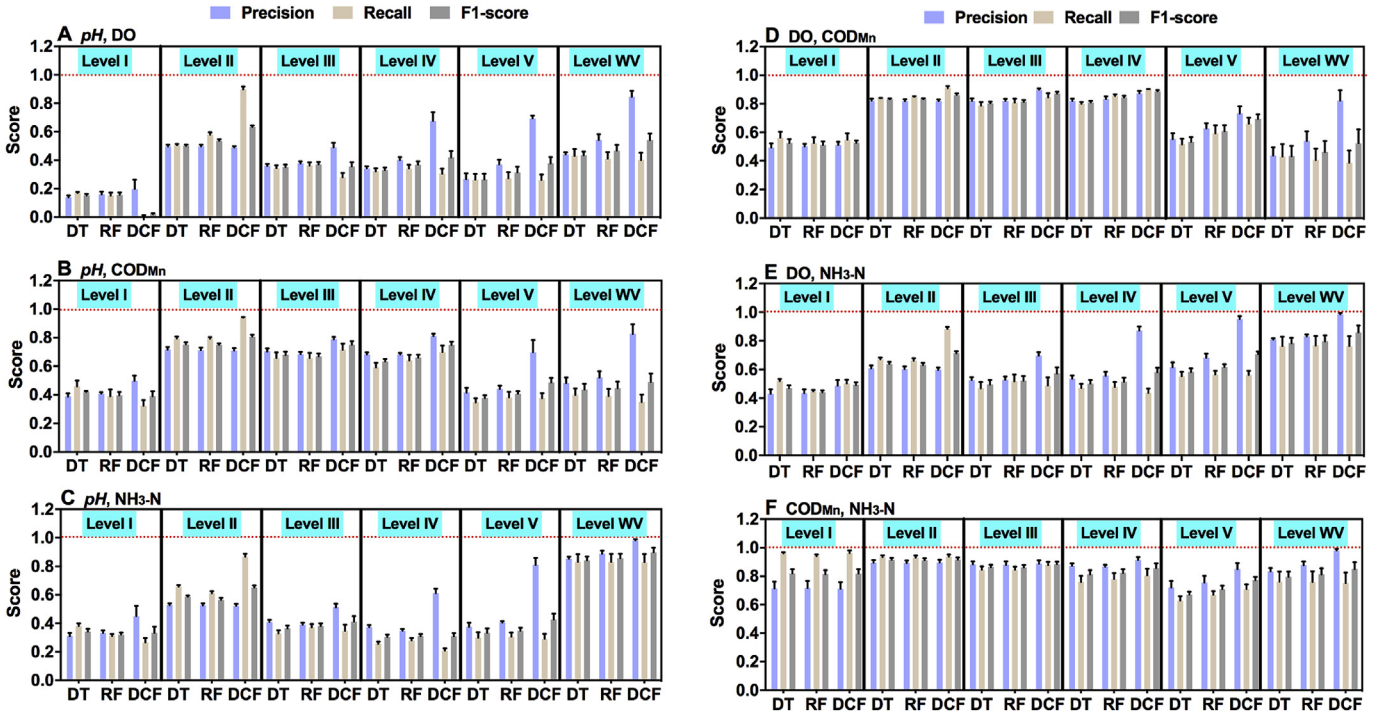

Fig. 8. The surface water quality prediction performance of 10 learning models in the validation process using 2 water parameters after training big data. (A) pH, DO; (B) pH, COD_{Mn}; (C) pH, NH_3—N; (D) DO, COD_{Mn}; (E) DO, NH_3—N; (F) COD_{Mn}, NH_3—N.

data. Notably, the greatest improvement of prediction performance in these models were identified when the training data increased from 1% (269 observations) to 10% (2689 observations). Interestingly, the minimum training subset (1%) were similar to the amount of training data set used in many previous studies, and the prediction performance of 10 models based the minimum training subset in this study were similar with the results in these previous studies (Walley and Džeroski, 1996). Consequently, the surface water quality prediction performance of learning models could be improved significantly by increasing the training data.

Furthermore, we observed diverse prediction performances by machine learning models based on the same big data set, especially those 7 traditional learning models. The underlying reason was probably the inherently different structures of those learning models. For 7 traditional machine learning models, DT and CRT showed better performance compared to other 5 learning models, especially DT with the highest precision, recall, F1-score, and weighted Fl-score for predicting all 6 Levels of water quality. This was because DT is a nonlinear classifier, which is optimal for dealing with the nonlinear classification problems including water quality prediction (Kim, 2016). Another reason was that DT does not require information on the prior probability of the feature distribution, and had better adaptability to sample imbalance issues (Liu et al., 2010). Imbalanced data sets can also have significant impact on the performance of LR and LDA models(Brown and Mues, 2012; Reed and Wu, 2013), which is confirmed in our result.

As a class of new learning models aims to improve over traditional models, ensemble learning models have been demonstrated to have better performance over many traditional learning models (Zhou, 2015). This is mostly because these ensemble learning models are optimized based on some traditional learning models, whilst retaining their advantages (Zhou, 2015). Therefore, RF, CRTF and DCF presented excellent prediction performance compared to all traditional learning models except for DT. DCF had the best performance for predicting all 6 Levels of water quality compared to other 9 machine learning models in this study. On the one hand, as a new kind of deep forest, DCF was an ensemble of the ensemble, which could further decrease the generalization error (Zhou and Feng, 2018). On the other hand, DCF has a hierarchical structure, which plays a key role in improving performance by increasing the final voting accuracy (Zhou and Feng, 2018). Moreover, DCF has the advantage of dealing with data of all sizes, due to DT and RF were selected to construct DCF in this study, and the better water quality prediction performance of DT and RF also were observed in this study. The three learning models especially DCF identified as the best learning model for water quality monitoring also could be potentially applied in the field of other environmental medium quality monitoring such as air quality monitoring. Similar to water quality monitoring, the air quality also depended on the air quality parameters (i.e. nitrogen oxide, ozone, and particulate matter). Additionally, these air parameters also monitored by environmental protection agencies around the world, which means the dig data of air quality was also available. However, DCF was unable to learn directly from the big raw data, which need do some work of feature engineering.

The relationship between different water parameters for learning models especially ensemble models and final prediction performance were identified for the first time in this study. With the decrease in the number of water parameters used, the prediction performance of DT, RF and DCF begun to deteriorate. However, the prediction performance for all 6 Levels of water qualities of DT, RF, DCF trained by DO, COD_{Mn}, and NH_3—N still was acceptable. Similarly, the better water quality performance of the 2 traditional models (artificial neural networks and genetic programming) trained by the same 3 fundamental water parameters were also observed (Muttil and Chau, 2007). However, in our study the performance of DT, RF, and DCF trained by these 3 water parameters were significantly better compared with this pervious study, due to bigger amount of training data. Moreover, the prediction performance for all 6 Levels of water qualities of DT, RF, and DCF trained by any 2 of 4 water parameters were also evaluated in our study. Although the decreased performance of the 3 models trained by 2 water parameters were identified compared with that trained by 3 and 4 water parameters, the COD_{Mn}, and NH_3—N still could be selected for future timely water quality prediction, based on the relative higher weighed F1-score.

Consequently, DO, COD_{Mn} and NH_3—N, as key water parameters, identified and validated by DT, RF and DCF would be recommend for future water quality prediction and monitor using learning models, which would significantly reduce prediction costs. In other words, the government administration could precisely issue water quality warning using least water parameters and spare more time for environmental experts to improve the water quality. The 2-parameter scenario using COD_{Mn} and NH_3—N could be applicated in cases such as the need for rapid prediction of water quality, although corresponding performance was lower than that trained by DO, COD_{Mn} and NH_3—N combined. For examples, when facing the cases of sudden environmental pollution, the quickest possible prediction of the Level of surface water contamination using the minimum possible measurements would be highly desirable. Nevertheless, future works are still needed to focus on identifying and validating other potential key water parameters among the available water parameters (*i.e.* water temperature, total phosphorus, metal ion, toxic organic chemicals), as well as identifying the key air parameters for air quality monitoring using different learning models based on this work.

5. Conclusions

The main aims of this study were to evaluate whether available big data could improve the performance of machine learning models in the prediction of surface water quality, and identify the best models and key water parameters serving timely and precisely water quality monitoring. Here, the surface water quality prediction performance of 7 traditional and 3 ensemble learning models using big data were comprehensively compared, and the potential key water parameters were also identified and validated. Through this study, the major conclusions are follows:

(1) Available big data could improve the performance of both traditional and ensemble learning models in the prediction of surface water quality.
(2) Compared to other 7 learning models, DT, RF and DCF presented significantly better prediction performance for all six Levels of water quality defined by Chinese government.
(3) Two key water parameter sets (DO, COD_{Mn}, and NH3—N; COD_{Mn}, and NH3—N) were identified and validated by learning models.

To sum up, the three learning models with two key water parameter sets identified and validated by big data in this study should be recommended for future surface water quality monitoring, because they not only could provide timely and precisely environmental warning, but also could further increase prediction efficiency and decrease the prediction cost in future surface water quality monitoring.

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.

Acknowledgements

This work was financially supported by National Natural Science Foundation of China (No. 61806096, 61872190, 61403208), China Postdoctoral Science Foundation (No. 2019M653280), and Foundation of Nanjing University of Posts and Telecommunications (No. NY220016).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.watres.2019.115454.

References

Aeberhard, M., Rauch, S., Bahram, M., Tanzmeister, G., Thomas, J., Pilat, Y., Homm, F., Huber, W., Kaempchen, N., 2015. Experience, results and lessons learned from automated driving on Germany's highways. IEEE Trans. Parallel Distrib. Syst. 7 (1), 42–57. https://doi.org/10.1109/MITS.2014.2360306.

Alabaster, J.S., Lloyd, R.S., 2013. Water Quality Criteria for Freshwater Fish. Butterworth, Cambridge, UK.

Avila, R., Horn, B., Moriarty, E., Hodson, R., Moltchanova, E., 2018. Evaluating statistical model performance in water quality prediction. J. Environ. Manag. 206, 910–919. https://doi.org/10.1016/j.jenvman.2017.11.049.

Barboza, F., Kimura, H., Altman, E., 2017. Machine learning models and bankruptcy prediction. Expert Syst. Appl. 83, 405–417. https://doi.org/10.1016/j.eswa.2017.04.006.

Belgiu, M., Drăguţ, L., 2016. Random forest in remote sensing: a review of applications and future directions. ISPRS J. Photogrammetry 114, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011.

Boulesteix, A.L., Janitza, S., Kruppa, J., König, I.R., 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. WIREs Data Mini. Knowl. Discov. 2 (6), 493–507. https://doi.org/10.1002/widm.1072.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Brown, I., Mues, C., 2012. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Syst. Appl. 39 (3), 3446–3453. https://doi.org/10.1016/j.eswa.2011.09.033.

Chen, J., Li, K., Tang, Z., Bilal, K., Yu, S., Weng, C., Li, K., 2016. A parallel random forest algorithm for big data in a spark cloud computing environment. IEEE Trans. Parallel Distrib. 28 (4), 919–933. https://doi.org/10.1109/TPDS.2016.2603511.

Di, Z., Chang, M., Guo, P., 2019. Water quality evaluation of the Yangtze River in China using machine learning techniques and data monitoring on different time scales. Water 11 (2), 339. https://doi.org/10.3390/w11020339.

Dunbabin, M., Marques, L., 2012. Robots for environmental monitoring: significant advancements and applications. IEEE Robot. Autom. Mag. 19 (1), 24–39. https://doi.org/10.1109/MRA.2011.2181683.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J., 2019. A guide to deep learning in healthcare. Nat. Med. 25 (1), 24–29. https://doi.org/10.1038/s41591-018-0316-z.

Fan, Y., Dong, H., Jiang, Y., Pan, J., Fan, S., Gui, G., 2018. An intelligent water regimen monitoring system. Int. Conf. Commun. Signal Process. Syst. 829–835. https://doi.org/10.1007/978-981-13-6508-9_101.

Ghahramani, Z., 2015. Probabilistic machine learning and artificial intelligence. Nature 521 (7553), 452–459. https://doi.org/10.1038/nature14541.

Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random forests for land cover classification. Pattern Recognit. Lett. 27 (4), 294–300. https://doi.org/10.1016/j.patrec.2005.08.011.

Gurovich, Y., Hanani, Y., Bar, O., Nadav, G., Fleischer, N., Gelbman, D., Basel-Salmon, L., Krawitz, P.M., Kamphausen, S.B., Zenker, M., 2019. Identifying facial phenotypes of genetic disorders using deep learning. Nat. Med. 25 (1), 60–64. https://doi.org/10.1038/s41591-018-0279-0.

Kim, K., 2016. A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree. Pattern Recognit. 60, 157–163. https://doi.org/10.1016/j.patcog.2016.04.016.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. https://doi.org/10.1038/nature14539.

Liu, W., Chawla, S., Cieslak, D.A., Chawla, N.V., 2010. A robust decision tree algorithm for imbalanced data sets. In: Proceedings of the 2010 SIAM International Conference on Data Mining, pp. 766–777. https://doi.org/10.1137/1.9781611972801.67.

MEEPRC, 2002. National Environmental quality standards of surface water of China, GB3838-2002. Ministry of Ecology and Environment of the People's Republic of China. http://www.mee.gov.cn/home/ztbd/rdzl/yjsp/fgbz/201110/W020061027509896672057.pdf. Accessed Dec. 23, 2019.

Muttil, N., Chau, K.-W., 2007. Machine-learning paradigms for selecting ecologically significant input variables. Eng. Appl. Artif. Intell. 20 (6), 735–744. https://doi.org/10.1016/j.engappai.2006.11.016.

Park, Y., Cho, K.H., Park, J., Cha, S.M., Kim, J.H., 2015. Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. Sci. Total Environ. 502, 31–41. https://doi.org/10.1016/j.scitotenv.2014.09.005.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12 (10), 2825–2830. http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf. Accessed Dec. 23, 2019.

Qiu, X., Ren, Y., Suganthan, P.N., Amaratunga, G.A., 2017. Empirical mode decomposition based ensemble deep learning for load demand time series forecasting. Appl. Soft Comput. 54, 246–255. https://doi.org/10.1016/j.asoc.2017.01.015.

Reed, P., Wu, Y., 2013. Logistic regression for risk factor modelling in stuttering research. J. Fluen. Disord. 38 (2), 88–101. https://doi.org/10.1016/j.jfludis.2012.09.003.

Simeonov, V., Stratis, J., Samara, C., Zachariadis, G., Voutsa, D., Anthemidis, A., Sofoniou, M., Kouimtzis, T., 2003. Assessment of the surface water quality in Northern Greece. Water Res. 37 (17), 4119–4124. https://doi.org/10.1016/S0043-1354(03)00398-1.

Singh, K.P., Gupta, S., Rai, P., 2013. Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmos. Environ. 80, 426–437. https://doi.org/10.1016/j.atmosenv.2013.08.023.

Singh, P., Kaur, P.D., 2017. Review on data mining techniques for prediction of water quality. Int. J. Adv. Res. Comput. Sci. 8 (5), 396–401. https://www.ijarcs.info/index.php/Ijarcs/article/viewFile/3312/3343. Accessed Dec. 23, 2019.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinf. 9 (1), 307. https://doi.org/10.1186/1471-2105-9-307.

Sun, Y., Chen, Z., Wu, G., Wu, Q., Zhang, F., Niu, Z., Hu, H.Y., 2016. Characteristics of water quality of municipal wastewater treatment plants in China: implications for resources utilization and management. J. Clean. Prod. 131, 1–9. https://doi.org/10.1016/j.jclepro.2016.05.068.

Tan, G., Yan, J., Gao, C., Yang, S., 2012. Prediction of water quality time series data based on least squares support vector machine. Procedia Eng. 31, 1194–1199. https://doi.org/10.1016/j.proeng.2012.01.1162.

Vörösmarty, C.J., McIntyre, P.B., Gessner, M.O., Dudgeon, D., Prusevich, A., Green, P., Glidden, S., Bunn, S.E., Sullivan, C.A., Liermann, C.R., 2010. Global threats to human water security and river biodiversity. Nature 467 (7315), 555–561. https://doi.org/10.1038/nature09440.

Walley, W., Džeroski, S., 1996. Biological monitoring: a comparison between Bayesian, neural and machine learning methods of water quality classification. Environ. Softw Syst. 229–240 https://doi.org/10.1007/978-0-387-34951-0_20.

Yu, P.S., Yang, T.C., Chen, S.Y., Kuo, C.M., Tseng, H.W., 2017. Comparison of random forests and support vector machine for real-time radar-derived rainfall forecasting. J. Hydrol. 552, 92–104. https://doi.org/10.1016/j.jhydrol.2017.06.020.

Zhang, C., Ma, Y., 2012. Ensemble Machine Learning: Methods and Applications. Springer, New York, USA.

Zhang, J., Mauzerall, D.L., Zhu, T., Liang, S., Ezzati, M., Remais, J.V., 2010. Environmental health in China: progress towards clean air and safe water. The Lancet 375 (9720), 1110–1119. https://doi.org/10.1016/S0140-6736(10)60062-1.

Zhou, Z., Feng, J., 2018. Deep forest. Natl. Sci. Rev. 6 (1), 74–86. https://doi.org/10.1093/nsr/nwy108.

Zhou, Z.-H., 2015. Ensemble learning. In: Encyclop. biomet, pp. 411–416. https://doi.org/10.1007/978-1-4899-7488-4_293.

Zhu, X., Vondrick, C., Ramanan, D., Fowlkes, C.C., 2012. Do we need more training data or better models for object detection?. In: Proceedings of the 23rd British Machine Vision Conference, pp. 1–11.