Natl. Acad. Sci. Lett. (May–June 2023) 46(3):203–207
https://doi.org/10.1007/s40009-023-01213-2

SHORT COMMUNICATION

Accurate Dissolved Oxygen Prediction for Aquaculture Using Stacked Ensemble Machine Learning Model

Rasheed Abdul Haq Kozhiparamban[1] · **P. Swetha**[1] · **V. P. Harigovindan**[1]

Received: 12 June 2022 / Revised: 24 October 2022 / Accepted: 4 January 2023 / Published online: 11 February 2023

Abstract Dissolved oxygen (DO) is the most vital water quality parameter that directly indicates the survival of aquatic life. Therefore, accurate DO prediction is essential for aquaculture water quality management for sustainable and profitable aquaculture production. Machine learning (ML) models have been successfully employed for water quality prediction. However, DO undergoes dynamic changes, which are nonlinear and complex, making accurate prediction of DO using conventional statistical methods and ML models a challenging task. To resolve this in this work, we propose a stacked ensemble ML model combining three different ML models as base learners and one ML model as a meta-learner to improve the DO prediction accuracy. The effectiveness of the stacked ensemble ML model has been evaluated using two different water quality datasets. The experimental results show that the stacked ensemble ML model achieves significant accuracy improvement compared with standalone ML models.

Keywords Aquaculture · Dissolved oxygen · Machine learning · Stacked ensemble model · Water quality prediction

Aquaculture farms explore practical approaches to reducing water consumption due to climate changes and water scarcity. Aquaculture plays a crucial role in ensuring food security for the growing population [1]. Intensive aquaculture systems are the fish farms with high stocking density, which can efficiently utilise land and water resource [2]. Maintaining optimum water quality plays a significant role in making intensive aquaculture production profitable, and this requires accurate water quality prediction. The dissolved oxygen (DO) concentration must be kept at optimum levels to stabilise the water quality. Decomposition of organic materials can quickly reduce the DO level in the water in a few hours, resulting in fish mortality [3, 4]. Accurate DO prediction resolves the issue and is crucial for maintaining aquaculture water quality [5].

Accurate prediction of DO is challenging as variations in DO are nonlinear and complex. Classical time series forecasting methods like Autoregressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving Average (SARIMA), Seasonal Autoregressive Integrated Moving Average with Exogenous Regression (SARIMAX), Holt Winter's Exponential Smoothing (HWES), etc. [6, 7], have problems associated with DO prediction such as less prediction accuracy and low generalisation [8]. As intensive aquaculture requires high accuracy, a novel water quality prediction technique for forecasting water quality parameters based on stacking is proposed in this work. Recently ensemble learning approach has been used in various fields which require high prediction accuracy. Stacking is a popular ensemble strategy to produce accurate results compared to a single learning model [9, 10]. The authors of [11] show a significant improvement in the forecast of PM 2.5 concentration using stacked selective ensemble-backed predictor (SSEP).

P. Swetha and V. P. Harigovindan have contributed equally to this work.

✉ Rasheed Abdul Haq Kozhiparamban
rasheedabdulhaq@gmail.com

P. Swetha
siriswethavas@gmail.com

V. P. Harigovindan
hari@nitpy.ac.in

[1] Department of Electronics and Communication Engineering, National Institute of Technology Puducherry, Karaikal, Puducherry 609609, India

The authors of [11] show a significant improvement in the forecast of PM 2.5 concentration using stacked selective ensemble-backed predictor (SSEP). The model accuracy is enhanced by employing the stacking ensemble integrating the merits of multiple single forecasting models. For the weekly forecasting of the data, in [12] authors propose a novel method that combines four base forecasting models using the lasso regression stacking approach. From the results, it is observed that the integrated forecast of various heterogeneous models on an average improve the accuracy of forecasting over the individual models. For the sales time series forecasting, authors of [13] propose a stacking approach for building a regression ensemble of single models. The results show that the overall performance of predictive models for sales time series forecasting is improved by stacking. Differently from existing works for DO prediction, we propose stacked ensemble machine learning (ML) model to improve the accuracy of DO prediction in this work. To the best of the authors' knowledge, this is the first research work to propose and analyse the performance of the stacked ensemble ML model to predict DO for aquaculture.

The significant contributions of this work are as follows:

1. Three years of data (January 2016 to December 2018) is collected from aquaculture ponds located in Kerala under the Agency for Development of Aquaculture Kerala (ADAK).
2. We propose a stacked ensemble ML model by integrating the merits of single forecasting models to achieve improved DO prediction accuracy for aquaculture.
3. We have considered the performance of the seven regression methods: support vector regression (SVR), random forest (RF), light gradient boosting machine (LGBM), elastic net (ENet), gradient boosting (GB), kernel ridge regression (KRR), and K-nearest neighbour (KNN). After considering the various possible combinations based on the optimal performance, we have chosen three different regression models as the base learners and one regression model as the meta-learner to implement the stacked ensemble ML model.
4. The performance of the stacking ensemble model is compared with standalone regression models using two different water quality datasets. The datasets used in this work are the water quality dataset collected from ADAK and a publically available dataset [8], which is collected from the marine aquaculture base in Xincun Town, LingShui County, Hainan Province, China. Results show that the stacking model significantly improves the accuracy of DO prediction compared to the standalone models.

Ensemble learning is a method where we train multiple models to solve a problem and are combined them to get more accurate results. Here the standalone performance of these single forecasting models called base learners need not be exceptionally good. The idea is that when weak base learners are combined with the right ensemble method, we get a more reliable and robust model. Three major approaches for combining the base learners in ensemble learning are bagging, boosting, and stacking. In this work, we have used stacking as the meta-algorithm to combine the weak base learners to provide highly accurate DO prediction for aquaculture.

Stacking is one ensemble concept for combining the predictions of base learners using a meta-learner. The data is divided into train and test sets. The input data is given directly to the base learners, the base learners predictions are provided as a new set of features to train the meta-model, and the predictions from the meta-model are the final prediction [14].

The main motive of introducing the stacking model is to keep down the generalisation error. Stacking typically provides better predictive performance compared with any single model. When the number of models used as base learners increases, we get better accuracy and improved generalisation. But the training time will also increase with the number of models in base learners. The performance of ensemble learning is best when we use the right combination of existing models as base learners. Meta-learner has the capability of learning to combine the best predictions of the base learners. The main concept of stacking is that the input data does not explicitly leak to the meta-model. To avoid data leakage, we use k-fold cross-validation at the base level. The aquaculture DO data collected from ADAK is divided into k folds, and for each iteration, $k-1$ folds get trained and make predictions on the untrained fold. The results of the k iterations are averaged to obtain a new set of feature for the next level. The process is replicated for all the base learners to produce the X_{meta} matrix to train the meta-model.

Stacking will produce significant results by preferably combining base learners using a meta-model because some models might work well in some parts of feature space, while other models might work well with others [15]. A significant improvement in the final predicted result using stacking can be achieved when a diverse set of models are used at different levels because different models follow different learning strategies. The diverse set of models at different levels will disagree with each other introducing a natural diversity that allows modelling various dynamic patterns in forecasting [16].

The proposed framework limits the number of models used at the base learner level to three. It reduces the training time, and no significant improvement is observed by increasing the number of base learners. The performance of the seven regression methods: support vector regression (SVR), random forest regression (RF) [17], light gradient boosting machine (LGBM) [18], elastic net regression

(ENet) [19], gradient boosting regression (GB), kernel ridge regression (KRR), and K-nearest neighbour regression (KNN) [20], is evaluated. We have three different regression models as the base learners and one regression model as the meta-learner to implement the stacked ensemble ML model. Various possible combinations are considered, and based on the optimal performance best combination is selected for the stacked ensemble ML model.

As illustrated in Fig. 1, the proposed method is a two-level stacking approach. The first level has three different regression techniques as base-level models, and the second level has one as meta-learner. We have considered all the possible combinations of seven regression methods: support vector regression (SVR), random forest (RF), light gradient boosting machine (LGBM), elastic net (ENet), and gradient boosting (GB). KRR, ENet, and LGBM as base learners and GB as a meta-model give optimal results compared to all other combinations. Hence for stacking of ML models, the regressors used in the base level are LGBM, ENET, and KRR, and the meta regressor is GB, and the cross-validation (CV) = 12.

The performance of the prediction models is evaluated using MAE (Mean Absolute Error), MSE (Mean Square Error), RMSE (Root Mean Squared Error) and MAPE (Mean Absolute Percentage Error), computed by the set of equations given below:

$$\begin{aligned} \text{MAE} &= \frac{1}{n}\Sigma_{i=1}^{n} \mid (A_i - Y_i) \mid \\ \text{MSE} &= \frac{1}{n}\Sigma_{i=1}^{n}(A_i - Y_i)^2 \\ \text{RMSE} &= \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(A_i - Y_i)^2} \\ \text{MAPE} &= \frac{1}{n}\Sigma_{i=1}^{n} \frac{|(A_i - Y_i)|}{A_i} \end{aligned} \tag{1}$$

where A_i is the actual value of i_{th} sample, Y_i is the predicted value of i_{th} sample, and n is the number of samples.

This research is to improve the prediction accuracy of aquaculture DO using a stacked ensemble ML model. These models are tested with two water quality datasets (ADAK and MAC). We use the 80% data from the dataset to train the model, and the remaining 20% is used to test the accuracy of the prediction results. The experimental environment is Microsoft Azure Virtual Machines with specifications: Inter(R) Xeon (R) 8272CL CPU @2.60GHz, 32 GB RAM, Windows 10 (64-bit) operating system, Visual studio code IDE, and we have implemented the neural network model using Python 3.9.6, Keras 2.6.0 and Tensorflow 2.6.0., Numpy 1.22.3, and Scikit-learn 1.0.2.

Table 1 summarises the comparison of MAE, MSE, RMSE and MAPE results of the stacked ensemble ML model with standalone ML models. Figure 2 plots the comparison of the predicted values using stacking and other ML models with the actual values for DO. It is clear from tables that stacking outperforms the standalone ML models for both datasets.

Figure 2a compares the predicted values using the stacked ensemble ML model and standalone ML models with the true data on the ADAK water quality dataset.

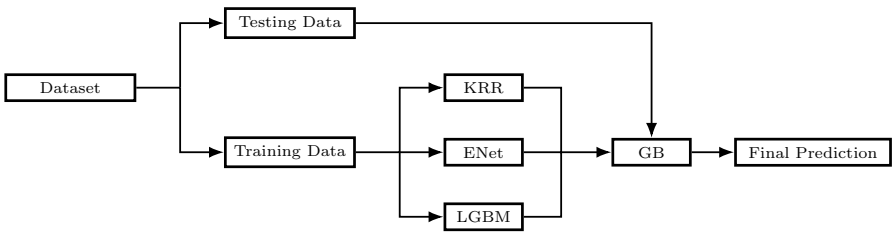
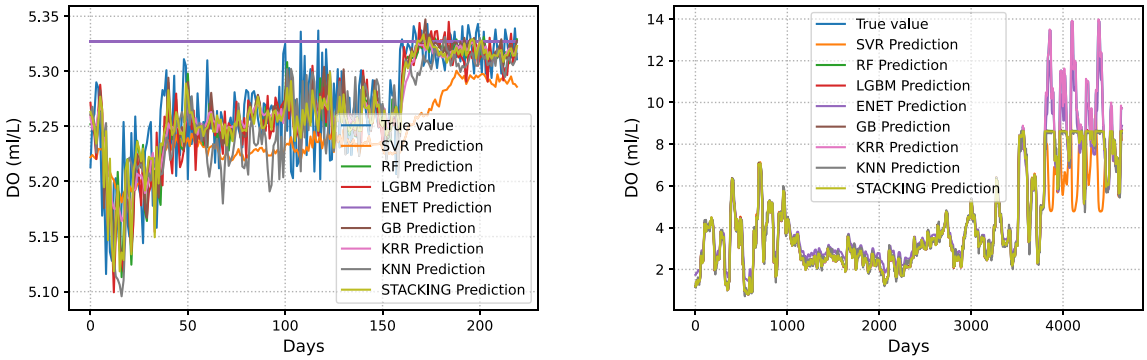


Fig. 1 Two-level stacking architecture of the proposed method

Table 1 Performance comparison of prediction accuracy of different ML models and stacked ML model for ADAK and MAC water quality datasets

Model	ADAK dataset				MAC datset			
	MAE	MSE	RMSE	MAPE	MAE	MSE	RMSE	MAPE
SVR	0.0377	0.0020	0.0450	0.0071	0.5123	2.6666	1.6330	0.0538
RF	0.0287	0.0014	0.0376	0.0055	0.2606	0.7525	0.8675	0.0279
LGBM	0.0296	0.0015	0.0385	0.0056	0.2705	0.7722	0.8787	0.0301
ENET	0.0611	0.0059	0.0767	0.0117	0.4123	0.3035	0.5509	0.1219
GB	0.0278	0.0013	0.0365	0.0053	0.2768	0.7760	0.8809	0.0318
KRR	0.0272	0.0013	0.0359	0.0052	0.2623	0.7825	0.8846	0.0272
KNN	0.0328	0.0018	0.0427	0.0062	0.3999	0.8721	0.9339	0.0657
Stacking	0.0255	0.0011	0.0332	0.0049	0.0176	0.0010	0.0319	0.0045



(a) ADAK dataset (b) MAC dataset

Fig. 2 Prediction performance comparison of different ML models and stacked ML model with true value for ADAK and MAC water quality dataset

For the ADAK water quality dataset, the prediction performance of all models is shown in Table 1. From the results, it is clear that the proposed stacked ensemble ML model (MAE = 0.0255 ml/L, MSE = 0.0011 ml/L, RMSE = 0.0332 ml/L and MAPE = 0.0049 ml/L) outperforms all ML models. The proposed model shows improvements of 6.16 %, 14.56%, 7.57 % and 6.19% beyond the best performing ML model KRR (MAE = 0.0272 ml/L, MSE = 0.0013 ml/L, RMSE = 0.0359 ml/L and MAPE = 0.0052 ml/L) on MAE, MSE, RMSE and MAPE, respectively.

Figure 2b compares the predicted values using the stacked ensemble ML model and standalone ML models with the true data on the MAC water quality dataset. For the MAC water quality dataset, the prediction performance of all models is shown in Table 1. From the results, it is clear that the proposed stacked ensemble ML model (MAE = 0.0176 ml/L, MSE = 0.0010 ml/L, RMSE = 0.0319 ml/L and MAPE = 0.0045 ml/L) outperforms all ML models. The proposed model shows improvements of 93.25 %, 99.87 %, 96.33 %, 83.74 % beyond the best performing ML model RF (MAE = 0.2606 ml/L, MSE = 0.7525 ml/L, RMSE = 0.8675 ml/L and MAPE = 0.0279 ml/L) on MAE, MSE, RMSE and MAPE, respectively.

In this research work, we have proposed the stacking ensemble regression model to improve the accuracy of the aquaculture DO prediction. We have considered the performance of the seven regression methods. Based on the optimal performance, we have chosen three different regression models (KRR, ENet and LGBM) as the base learners and the GB regression model as the meta-learner to implement stacking ensemble regression. These prediction models were trained and tested on two distinct datasets. We have compared the performance of the stacking ensemble ML model with standalone models in terms of MAE, MSE, RMSE and MAPE. Results show that the stacking model significantly improves prediction accuracy compared to the standalone ML models, offering a realistic solution for forecasting the water quality parameter DO in aquaculture.

References

1. Jennings S, Stentiford GD, Leocadio AM, Jeffery KR et al (2016) Aquatic food security: insights into challenges and solutions from an analysis of interactions between fisheries, aquaculture, food safety, human health, fish and human welfare, economy and environment. Fish Fish 17(4):893–938
2. Oddsson G (2020) A definition of aquaculture intensity based on production functions-the aquaculture production intensity scale (apis). Water 12(3):765
3. Ayele HS, Atlabachew M (2021) Review of characterization, factors, impacts, and solutions of lake eutrophication: lesson for lake Tana, Ethiopia. Environ Sci Pollut Res 28(12):14233–14252
4. Martos-Sitcha JA, Mancera JM, Prunet P, Magnoni LJ (2020) Editorial: welfare and stressors in fish: challenges facing aquaculture. Front Physiol 11:162
5. Boyd CE, Torrans EL, Tucker CS (2018) Dissolved oxygen and aeration in ictalurid catfish aquaculture. J World Aquac Soc 49(1):7–70
6. Nagaraj N, Mohan BR (2020) Intraday stock prediction based on deep neural network. Natl Acad Sci Lett 43(3):241–246
7. Mitra D, Paul RK (2021) Forecasting of price of rice in India using long-memory time-series model. Natl Acad Sci Lett 44(4):289–293
8. Liu J, Yu C, Hu Z, Zhao Y et al (2020) Accurate prediction scheme of water quality in smart mariculture with deep bi-s-sru learning network. IEEE Access 8:24784–24798
9. Buyrukoğlu S, Savaş S (2022) Stacked-based ensemble machine learning model for positioning footballer. Arab J Sci Eng. https://doi.org/10.1007/s13369-022-06857-8
10. Rahman LF, Marufuzzaman M, Alam L, et al (2022) Application of machine learning to investigate the impact of climatic variables on marine fish landings. Natl Acad Sci Lett 45:245–248. https://doi.org/10.1007/s40009-022-01110-0
11. Gu K, Xia Z, Qiao J (2020) Stacked selective ensemble for pm2.5 forecast. IEEE Trans Instrum Meas 69(3):660–671
12. Godahewa R, Bergmeir C, Webb GI, Montero-Manso P (2020) A strong baseline for weekly time series forecasting. CoRR, abs/2010.08158

13. Pavlyshenko B (2019) Machine-learning models for sales time series forecasting. Data 4(01):15
14. Pavlyshenko B (2018) Using stacking approaches for machine learning models. In: 2018 IEEE second international conference on data stream mining processing (DSMP), pp 255–258
15. Ribeiro MHDM, dos Santos Coelho L (2020) Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. Appl Soft Comput 86:105837
16. Dutta H (2009) Measuring diversity in regression ensembles. In: IICAI. Citeseer, vol 9, p 17p.
17. Cutler A, Cutler D, Stevens JR (2011) Random forests. Mach Learn 45(157–176):01
18. Alzamzami F, Hoda M, Saddik AE (2020) Light gradient boosting machine for general sentiment classification on short texts: a comparative evaluation. IEEE Access 8:101840–101858
19. Rapach DE, Zhou G (2020) Time-series and cross-sectional stock return forecasting: New machine learning methods. In: Machine learning for asset management: new developments and financial applications, pp 1–33
20. Taunk K, De S, Verma S, Swetapadma A (2019) A brief review of nearest neighbor algorithm for learning and classification. In: 2019 international conference on intelligent computing and control systems (ICCS), pp 1255–1260