WITHDRAWN: The Effect of Reduced Water Quality Variables on D issolved Oxygen Concentration Prediction Performance of Random Forest and Multilayer Perceptron Regressors

FARID HASSANBAKI GARABAGHI

fgarabaghi@gmail.com

Gazi University: Gazi Universitesi https://orcid.org/0000-0003-0244-0360

Semra Benzer

Gazi University: Gazi Universitesi

Recep Benzer

Baskent University: Baskent Universitesi

Research Article

Keywords: Regression, Feature Selection, Prediction, Random Forest, Multilayer Perceptron, Dissolved Oxygen

Posted Date: May 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1554196/v1

EDITORIAL NOTE:

The full text of this preprint has been withdrawn by the authors while they make corrections to the work. Therefore, the authors do not wish this work to be cited as a reference. Questions should be directed to the corresponding author.

Abstract

That life highly depends on oxygen on the blue planet is an undeniable fact. Not only for preserving the ecosystem as well as aquatic organisms' well-being, but sufficient amount of dissolved oxygen (DO) in water systems is also important from the sustainable drinking water resources point of view. Therefore, as important as monitoring DO concentrations in aquatic systems, accurate prediction of future levels of DO has become in the center of attention of many researchers as well as national and international agencies. One simple way to overcome this task is to get benefit from machine learning methods. This study was designed to evaluate the performance of random forest (RF) regressor against multilayer perceptron (MLP) as a popular model of artificial neural networks in predicting DO concentrations. More importantly, the effect of dimension reduction of the dataset by applying wrapper feature selection technique on the performance and accuracy of prediction of the regressors was also investigated. The results showed that RF regressor excelled MLP in performance with either of the datasets (the dataset of all parameters of interest and the dataset of reduced subset of features). Moreover, training the RF regressor with reduced subset of features positively affected the performance of the regressor, however, the accuracy of the MLP regressor declined considerably when it was trained with the dataset of reduced features. The results proved the positive impact of dimensionality reduction on the performance of RF regressor for predicting DO concentrations.

1. Introduction

That existence would not be possible without oxygen is an undeniable fact. Not being an exemption, aquatic life does highly depend on the level of dissolved oxygen (DO) in water (Shaghaghi et al. 2020). Moreover, once some elements such as nitrogen, sulfur, phosphorous, etc. might have toxic effects on aquatic organisms, combined with DO, essential compounds for the survival of aquatic organisms are formed (Araoye 2009). In the last decades, factors (i.e., uncontrolled urbanization, industrial refuse, agricultural activities, etc.) related to excessive population growth have led to surface water pollution and consequently low levels of DO which has significantly constrained the living conditions for aquatic organisms (Ouma et al. 2020). Hence, being one of the essential parameters reflecting the quality of water (Ma et al. 2018), monitoring DO levels is of a paramount importance for environmental scientists to decide about the health of water systems. Apart from this, since frequent monitoring of DO is highly time consuming, today, scientists and decision-making agencies tend to make use of models to predict DO concentration based on other water quality parameters or estimate the DO concentration solely based on previous DO values over time and no other water quality parameters (Heddam and Kisi 2017). To reach this goal, numerous attempts with and without machine learning approach have been conducted and the regression performance of many algorithms for DO concentration prediction have been evaluated. For instance, Heddam and Kisi (2017) evaluated the DO concentration prediction accuracy of Multilayer Linear Regression (MLR), Multilayer Perceptron (MLP), and Extreme Learning Machine (ELM) algorithms with and without water quality parameters as indicators. As a result, generally, EML models outperformed the other two regression models, yet MLP showed better performance over ELM specially where no water quality parameters were involved in the process of prediction. Elsewhere, Li et al. (2017) objected two

goals in their investigation in regard with the prediction of DO concentration which were firstly the comparison of three machine learning models (i.e., MLR, back propagation neural network (BPNN), and support vector machine (SVM)) and secondly the effect of an optimizer on the performance of the regressors. Not only they reported the positive effect of the optimizer algorithm on the performance of the regressors, but the superiority of the SVR algorithm, combined with the optimizer, over the other two algorithms was reported as well. Again, for hourly prediction of DO concentration, Kisi et al. (2020) evaluated the performance of Bayesian Model Averaging (BMA) with five other machine learning algorithms (i.e., ELM, ANNs, adaptive neuro-fuzzy inference system (ANFIS), classification and regression tree (CART), and MLR). Resultingly, the superiority of BMA over the other five algorithms was reported. Ouma et al. (2020) compared the DO concentration regression performance of Feedforward Neural Network (FNN) model with MLR model in which the superiority of FNN over MLR model was reported. Validating the result of the previously mentioned study, elsewhere, Zhu and Haddam (2020) compared the DO concentration prediction performance of Multilayer Perceptron Neural Network (MLPNN) Model, which is a fully-connected Feedforward Neural Network model, with EML as a result of which the slight excellence of MLPNN over EML model was reported. Valera et al. (2020) conducted an investigation towards performance evaluation of SVR and Random Forest Regressor (RFR) in prediction of DO. They reported a slight merit of RF regressor over SVR. Ahmed and Lin (2021) evaluated the performance of Quantile Regression Forest (QRF) algorithm in prediction of DO and eventually compared its performance with MLPNN. Reportedly, the performance of QRF model surpassed MLPNN. Dehghani et al. (2021) assessed not only the performance of SVR in prediction of DO concentration, but they also did evaluate the effect of four optimizers on the performance of the model. It is reported that the optimizers could significantly increase the performance of the model.

Obviously, many scientists have conducted investigations to introduce a noteworthy machine learning algorithm capable of predicting DO with higher accuracy. Some of them also reported the difference between the accuracy of prediction with and without water quality parameters. Others, to increase the accuracy rate of the algorithms, have used some techniques such as employment of optimizers. Yet a lack in the literature is quite tangible which is the investigation of the most effective variables in the process of prediction of DO as well as the impact of reduced variables on the accuracy of the models.

Furthermore, catching a glance at the literature review above, which is the abstract of similar studies, and comparing the results in detail, ANN models by and large have been suggested as promising methods in prediction studies. The second lack in the literature is that ANNs have not challenged with ensembled algorithms such as Random Forest Regressor.

Therefore, this study was aimed to follow two objectives. First of all, the performance of Random Forest Regressor was weighed against Multilayer Perceptron which is a model of ANNs. Secondly, the effect of reduced features on the performance of the models was evaluated.

2. Material And Method

2.1. Study area and the dataset

One of the most prone to contamination basins in Turkey with the area of 2,600,967 ha (Büyük Menderes Basin) was aimed to be investigated in this study. Most of the contamination load of the basin comes from industrial activities such as agriculture, livestock, food and tourism.

In this study, the water quality data of 8 stations along Büyük Menderes River which is the main stream in the basin was provided by the General Directorate of State Hydraulic Works in Turkey. The dataset contained the water quality data from 2004–2014. The water quality measurement intervals were two-month.

A dataset of 528 instances and 13 parameters (i.e., year, month, station, temperature, pH, electrical conductivity, chemical oxygen demand, biological oxygen demand, ammonium nitrogen, nitrite, nitrate, phosphorous and dissolved oxygen) was generated after the process of data integration (Kumar and Manjula 2012) to create a uniform dataset.

2.2. Machine learning approach

Derived from artificial intelligence concept, machine learning has become a trend among science communities to facilitate and automate complicated tasks through experience (Jordan and Mitchell 2015). Simply, machine learning algorithms are used to handle data problems and since there is no one-fit-for-all algorithm to solve all problems, an algorithm must be chosen through experiments to be suggested for handling a particular problem (Mahesh 2020). Therefore, in this study, the performance of two machine learning algorithms-one which has been widely used and suggested by many researchers (multilayer perceptron) for DO prediction and the other which has been disesteemed in dealing with such problems (random forest)-in dealing with prediction of DO concentrations have been evaluated.

2.2.1. Artificial neural networks (ANNs)

Imitating mankind brains pattern in learning, artificial neural networks (ANNs) were generated to process complex information to make logical decisions. Artificial neural networks consist of three main layers known as input layer through which the data is entered the network and is processed and computed in the hidden layer to make a logical decision in the output layer. These layers are interconnected to each other through nodes mocking the human brains neuron system. Fundamentally, the relation between input and output of the system is sigmoid, however, depending on the data and the purpose of the study, these functions may vary. What's more, basically, feeding one logical decision to the next one, the connectivity pattern between the nodes is feed-forward in ANNs (Hopfield 1988). In this study, experimentally-based on the best achieved results- a model of multilayer perceptron (MLP) comprises of two substrates of hidden layer with 5 and 3 nodes respectively, at learning rate of 0.03 with the momentum of 0.2 was generated and the best results achieved by this model are discussed in the following section(s).

2.2.2. Random forest (RF)

Random forest which is an ensemble learning model working with bagging technique consist of several base models which are generally decision tree models (Fig. 1) (Breiman 2001). This model is promised to solve classification and regression problems specially where the dataset consists immense number of outliers as well as noise (Benzer et al. 2022). In this model, each tree model is fed by a random sample of dataset with equal distribution among all the trees in the forest. Finally, the results achieved by the tree models are aggregated and averaged to become the output of the forest (Breiman 2001). Although the performance of RF model in classification problems has been investigated by many researchers, the performance of the model in dealing with regression problems specially prediction of DO can be considered further. Therefore, in this study RF was employed to predict DO concentrations.

2.3. Feature selection

Keeping the prediction accuracy high along with eliminating irrelevant variables as well as the noise, the objective of feature selection is to generate a subset of features from the main dataset representing the whole input data efficiently (Chandrashekar and Sahin 2014). Feature selection can make an algorithm computationally cost-effective by reducing the dimension of the dataset resulting in shrinking the storage space which together lead to shorter training time and in some cases may increase the accuracy of the prediction or classification (Danaei Mehr and Polat 2021). Feature selection methods are categorized into three main groups known as (i) filter method, (ii) wrapper method, and (iii) embedded method. In this study, many subcategories of these three main groups have been tested to select the most efficient subset of features and because the subset of features selected by wrapper method achieved the best results by the selected machine learning algorithms for the prediction of DO, the other tested feature selection methods and the results are not mentioned in this article.

2.3.1. Wrapper feature selector

Wrapping around the induction algorithm (Fig. 2) and evaluating the features using the accuracy and error rate of the classification or prediction process to select the best and most distinguished features, wrapper approach or so-called close-loop method diminish the estimation error of a particular induced classifier (Kohavi and John 1997; Zebari et al. 2020). Since this method selects the most optimal features and eliminates irrelevant attributes, it attains better performance in comparison with other feature selectors (Zebari et al. 2020). In this study, wrapper feature selection method with K-star lazy algorithm as the induction algorithm with forward direction was applied to generate the most efficient subset of features for the process of prediction of dissolved oxygen concentrations.

2.4. Evaluation metrics

In this section evaluation metrics, which investigate efficiency of the regression models are described below.

2.4.1. Pearson correlation coefficient

The Pearson correlation coefficient, which is an influence indicator of input values on the dependent values, measures the correlation degrees among multiple variables. The value of correlation degree is between [-1,1] in which 1 indicates the strong correlation and − 1 indicates weak correlation among variables. As a result, the lower absolute value of correlation degree is, the weaker correlation is and vice versa. The Pearson correlation coefficient formula is depicted by Eq. 1 Where,

E(X) determine the expected value of X and cov (X, Y) and σ_X σ_Y define covariance of X, Y and standard deviation of X and Y respectively (Lord et al. 2021).

$$p_{X,Y} = \frac{\operatorname{cov}(\mathrm{X},\mathrm{Y})}{\sigma_X \sigma_Y} = \frac{\mathrm{E}\left((\mathrm{X}-\mu_X)(\mathrm{Y}-\mu_X)\right)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}(\mathrm{XY}) - \mathrm{E}(\mathrm{X})\,\mathrm{E}(\mathrm{Y})}{\sqrt{E\left(X^2\right) - E^2(X)}\sqrt{B\left(Y^2\right) - E^2(Y)}}$$

1

2.4.2. Mean Absolute Error

Mean absolute error (MAE) is a vector-to-vector regression evaluation method which measures that to what extent the predicted and estimated values are close to actual values. Hence, the lower value of MAE, the lower differences between predictions and the actual values and vice versa. Eq. 2 is the MAE equation in which,

n indicates the number of test or training set and A_t and F_t are actual and predicted values (Qiao et al. 2019).

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |A_t - F_t|$$

2

2.4.3. Root Mean Square Error

Root Mean Square Error (RMSE) which is known as root mean standard deviation, is used to assess the performance of the predictions. Therefore, it measures the distance (residual) between predictions and actual values, calculates the mean value of residuals and computes the square root of mean values. RMSE is defined as below (Eq. 3) in which,

N represents number of values, y_i and $\hat{y}_i$ indicate the i^{th} value and its related prediction (Rustam et al. 2020).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \left\| y_{i-\hat{y}_i} \right\|^2}{N}}$$

3

3. Results And Discussion

In this study, among 13 parameters of interest (i.e., year, month, station, temperature (T), pH, electrical conductivity (EC), chemical oxygen demand (COD), biological oxygen demand (BOD), ammonium nitrogen (NH_4-N), nitrite (NO_2), nitrate (NO_3), phosphorous (PO_4) and dissolved oxygen (DO)) which were chosen to assist the ML algorithms in prediction of DO, 9 parameters were selected by the wrapper feature selector to reduce the dimension of the dataset. The plot matrix of the selected variables is illustrated in Fig. 3. Although the eliminated variables were in close association with DO concentration, the wrapper feature selector found them the cause of noise (Chandrashekar and Sahin 2014) in the dataset and the most related and distinct variables associated with DO was found to be year, month, station, T, pH, EC, COD, BOD5, and NO_3.

Discussed in the previous section (concept of feature selection methods), the mechanism which a feature selection algorithm works with is to eliminate irrelevant variable as well as keeping the most distinct variables in order to reduce the noise with the objective of increasing the performance of the classifier or the predictor. The result of the used wrapper feature selector shows that the algorithm tended to keep only one nitrogen source to assist the predictors. What's interesting is that the feature selector has chosen nitrate over nitrite to keep in the generated subset of features. Along with ammonium nitrogen, nitrite and nitrate, phosphorous which is the main source of nutrient contamination in ground waters causing algal blooming and consequently significantly diminish the DO concentrations (Rozemeijer and Broers 2007; Varol and Şen 2012) which demonstrates its close association with DO levels in water systems was eliminated from the set of variables of interest. Furthermore, despite the mechanism of eliminating identical variables in order to reduce the noise as well as dimension, COD and BOD5 together were kept in the generated subset of features by the wrapper feature selector.

Validating the variables selected by the wrapper feature selector, besides observation information (i.e., year, month, and station), the stratification indicators such as T and pH and mineral budget indicators such as EC (Ouma et al. 2020) have been found to be essential parameters in prediction of DO concentration as many researchers have chosen them as input water quality variables based on their expertise (Heddam and Kisi 2017; Keshtegar and Heddam 2018; Csábrági et al. 2019; Ouma et al. 2020; Zhu and Heddam 2020, Dehghani et al. 2021). Other than the mentioned variables, some researchers have also included parameters such as COD (Zhu and Heddam 2020). Yet the effect of parameters such as NO3 and BOD5 which are in a close association with DO levels in water systems needs further considerations in such studies. The effect of the selected parameters by the wrapper feature selector on the performance of the predictors will be discussed in the following section(s).

Table 1, shows the results of the metrics used to evaluate the performance of each ML model applied for the prediction of the DO concentration in the target basin. Evidently, RF regressor showed better performance over MLP in terms of all metrics, yet the effect of reduced variables on the performance of the regressors was considerable. While the performance of RF regressor slightly increased when 9 variables were included in the process of prediction, the performance of MLP declined significantly by

considering the reduced subset of features. This can demonstrate the adverse effect of applying reduced subset of features on ANN algorithms as they tend to observe as much as there is variables in the training phase in order to make logical decision in the end and related variables don't have serious adverse effect on the performance of ANN models. In other word, it can be concluded that ANN models are less prone to noise than tree-based models such as random forest.

Csábrági et al. (2019) reported in their work in which they have compared the performance of differently optimized ANN models that the performance of the ANN model significantly increases with a larger dataset. Ouma et al. (2020) also reported in their article in which they have compared the performance of an ANN model with MLR statistical model that the ANN model's performance increase when all the variables of interest are included in the process of DO prediction.

Catching another glimpse of Table 1, although the correlation coefficient of the MLP regressor dropped down by two percent, MAE, which indicates the intimacy of the prediction, hasn't increased significantly. Therefore, it can be concluded that the general proximity of estimations from the actual DO concentrations are virtually the same. However, despite slight increase in the correlation coefficient of the RF regressor with reduced subset of features, MAE as well as RMSE considerably declined which demonstrate the promotion of the estimations by the algorithm when the noise and dimension drops down.

Table 1
Performance evaluation metrics of the regressors

Model	Variables	Pearson Correlation Coefficient	Mean Absolute Error	Root Mean Square Error
RF	Whole	0.7958	0.9188	1.3093
MLP	Whole	0.7586	1.2424	1.7291
RF	Reduced	0.8052	0.8911	1.2805
MLP	Reduced	0.7367	1.2495	1.7525

Fig. 4, depicts fitness of the predicted values of DO made by the RF regressor by considering the whole variables of interest to the actual values of DO. Obviously, for DO concentrations between 5 to 9 mg/l the predicted values by the algorithm fit well to the actual values of DO. Moreover, values lower and higher than this range couldn't be perfectly estimated by the algorithm.

Diving deep into the RMSE of the RF model trained with the whole variables of interest (Fig. 5), the estimation error deviation (EED) of the algorithm is less (somewhere between -1 to 1.3) for 72.91% of the values predicted in the testing phase which can be considered a promising general error deviation for the model trained by the aforementioned set of features because although the remaining values' (27.09%)

error deviation exceed this range, only 7.38% of the predicted values were in unacceptable estimation error deviation range (2.34<EED<-2.22).

Compared to the fitness of the actual vs predicted scatter plot of the RF regressor when all features of interest were considered, the predicted values of DO with reduced subset of features look more fit to the actual values in Fig. 6. Alike the results of Fig. 4, it seems the RF regressor, when reduced subset of features supervised it, could remarkably predict the values of DO between concentrations 5 to 9 mg/l.

However, the estimation error deviation histogram plot of the model with reduced subset of features (Fig. 7) shows and demonstrates the positive effect of the noise cancelation and dimension reduction of the dataset on the prediction ability of the regressor as the regressor could predict 73.86% of the instances with the remarkable estimation error deviation range of -1 to 1.26. Moreover, only 6.81% of the instances predicted in an unacceptable estimation error deviation range (-2.1 < EED < 2.38).

Conversely, MLP algorithm demonstrated that its prediction ability highly depends on the number of variables in the training phase (Table 1). In other word, the more variables are included in the training phase, the more accurately MLP algorithm can predict the values in the testing phase.

Figure 8 and 9, respectively, depict the scatter plot of the actual versus predicted DO values predicted by the MLP regressor with the whole variables of interest and the reduced subset of features. Comparatively, the fitness of the plot is remarkable when all the variables of interest were included in the training phase. These two figures, undoubtedly demonstrate inevitable dependability of the MLP algorithm to the training from more sources rather than limited number of features which is contrary to the mechanism and results of tree-based algorithms as they are susceptible to excessive number of features for prediction of a target value (Kohavi and John 1997). Just alike RF regressor's results, the most accurate DO value prediction bond with the MLP algorithms is 5 to 6 mg DO/l.

As for the error estimation deviation of the model (MLP) when all variables of interest are included in the training phase (Fig. 10), 71.51% of the predicted DO values were in the outstanding EED range of -1.44 to 1.35 whereas 77.21% of the predicted DO values by the algorithm were in this range when reduced subset of features were included in the training phase (Fig. 11). This aligns with the EED results of the RF model as when the reduced subset of features was included in the training phase, both models' EED declined significantly which proves the remarkable effect of noise cancelation as well as dimension reduction on the accuracy of estimation of the both regressors.

The performance of ANN models as well as tree-based models including random forest as a robust tree-based algorithm working with ensemble learning technique have been investigated in many research lines to propose promising models to solve classification and/or regression problems.

Results of the current study showed that RF is a promising regressor in predicting DO concentrations using other water quality parameters. Despite satisfying results, MLP algorithm exhibited weaker

performance against RF. Furthermore, the effect of feature selection on the performance of the regressors is undeniable.

Over years and paying serious efforts, the superiority of RF and its derivatives over many other ML models in solving regression problems have been reported as well. For instance, Singh at al. (2013) reported the superiority of RF algorithm over simple decision tree as well as SVR benchmark model in predicting air quality parameters. Wang et al. (2016) evaluated the performance of RF and Random Bits Forest (RBF) which is a method derived from the original concept of RF with seven other ML algorithms in solving a regression problem. Resultingly, the superiority of RF and RBF models over others was reported. Maheshwari and Lamba (2019) investigated the performance of 6 regressors including RF and MLP in solving an air quality regression problem as a result of which, despite an outstanding performance exhibition by MLP, random forest regressor outperformed the other four ML regressors in accuracy. It can also be concluded from the results of this study that the MLP regressor is a robust algorithm in solving such regression problems. Ghorbani et al. (2016) reported the regression performance excellence of MLP over SVR model in predicting river flow. In a different study, Heidari et al. (2016) investigated the regression ability of the MLP algorithm in prediction of the viscosity of nanofluids and they achieved an outstanding result applying the MLP algorithm in solving such regression problem which demonstrates its exceptional capability of prediction. However, that doesn't mean the feebleness of other regressors compared to the MLP. For example, Amid and Gundoshmian (2016) investigated the performance of three regressors including MLP in prediction of the output energies of boilers. As a result, MLP showed the weakest performance among the group of three regressors. There are other examples in the literature as results of which MLP wan not found to be an effective algorithm to solve regression problems (Ghritlahre and Prasad 2018). Therefore, firstly due to this algorithm's popularity and of course its simplicity of application, its regression performance has to be investigated further through different scenarios.

Recently, feature selection has attracted many researchers' attention in different research domains due to its capability of making the implementation of regression and/or classification cost-effective. Reduced subset of features is generated from the main dataset not only to reduce the dimension of the dataset which decrease the extra computations and consequently make a model feasible, but also to improve the performance of prediction. However, not always reduced subset of features rocket up the performance but it might impact the performance of some ML regressors adversely.

Masmoudi et al. (2020) reported the improvement in prediction of the regressor when feature importance framework is applied to reduce the dimension of the dataset. Castangia et al. (2021) reported the positive impact of the reduced subset of feature on the performance of five ML regressors including RF regressor.

The results of our study also proved the positive impact of training the RF regressor with reduced subset of features for predicting DO concentrations. Yet the adverse effect of the reduced features on the performance of MLP regressor for prediction of DO is undeniable. However, there are studies in the literature in which the effect of reduced subset of features on the performance of MLP regressor was investigated and the results are in contradiction with the results of this study. For example, Hossain et al.

(2013) reported the improvement in the accuracy of MLP regressor for the prediction of solar power using the reduced subset of features.

Considering the benefits of feature selection and reduced subset of features, sacrificing a bit of accuracy for considerable declining computations doesn't seem much irrational. However, when a potent regressor such as RF can achieve brilliant results with or without feature selection, it won't be cost-effective to chase ineffective algorithm for prediction of DO concentration in this study. Yet the robustness and accuracy of the MLP regressor can be investigated through different scenarios and different datasets.

4. Conclusion

The objective of this study was to investigate the performance of random forest (RF) and multilayer perceptron (MLP) regressors in predicting DO concentrations in water with and without feature selection. Although both regressors performed well in predicting DO values, RF excelled MLP when trained with and without reduced features. Moreover, reduced features showed positive impact on the performance of RF regressor, however, the performance of MLP regressor declined significantly by applying the reduced subset of features. It can be concluded that the MLP regressor is quite susceptible to the number of parameters in training phase for predicting DO. In other word, the more parameters are included for predicting DO concentrations in the training phase, the more accurately MLP regressor can predict the values. All of all, it can be concluded that for this particular dataset which was used in this study, RF is suggested as an effective regressor for predicting DO concentrations in Buyuk Menderes Basin specially when reduced subset of features is used in the training phase because random forest regressor showed great sensitivity to the dimension of the dataset.

Declarations

Acknowledgement

Special thanks to the General Directorate of State Hydraulic Works, Department of Survey, Planning and Allocations for provision of the water quality data of Büyük Menderes Basin and all the supports thereafter.

Declaration of preprint submission

We hereby declare that the preprint of the presented manuscript has been submitted in Research Square not prior to submission on Environmental Science and Pollution Research.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions

Formal analysis, visualization, investigation and software were performed by **Farid Hassanbaki Garabaghi.** Conceptualization and methodology were performed by **Semra Benzer** and **Recep Benzer.** The first draft of the manuscript was written by **Farid Hassanbaki Garabaghi,** and **Semra Benzer** and **Recep Benzer** reviewed, commented and edited the previous versions of the manuscript.

Ethical approval

This study did not involve human participants, human material, or human data so it does not need ethical approval document.

Consent to participate

All authors have agreed to participate to this study

Consent for publication

All authors have agreed to be coauthors of this manuscript.

Availability of data

Not applicable

References

1. Amid S, Gundoshmian TM (2016) Prediction of Output Energies for Broiler Production Using Linear Regression, ANN (MLP, RBF), and ANFIS Models. Environ Prog Sustain Energy 36(2):577–585. https://doi.org/10.1002/ep.12448
2. Ahmed MH, Lin LS (2021) Dissolved oxygen concentration predictions for running waters with different land use land cover using a quantile regression forest machine learning technique. J Hydrol 597:126213. https://doi.org/10.1016/j.jhydrol.2021.126213
3. Araoye PA (2009) The seasonal variation of pH and dissolved oxygen (DO_2) concentration in Asa Lake Ilorin, Nigeria. Int J Phys Sci 4(5):271–274. https://doi.org/10.5897/IJPS.9000580
4. Benzer S, Garabaghi FH, Benzer R, Mehr HD (2022) Investigation of some machine learning algorithms in fish age classification. Fish Res 245:106151. https://doi.org/10.1016/j.fishres.2021.106151
5. Breiman L (2001) Random Forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324
6. Castangia M, Aliberti A, Bottaccioli L, Macii E, Patti E (2021) A compound of feature selection techniques to improve solar radiation forecasting. Expert Syst Appl 178:114979. https://doi.org/10.1016/j.eswa.2021.114979

7. Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40:16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024
8. Csábrági A, Molnár S, Tanos P, Kovács J, Molnár M, Szabó I, Hatvanid IG (2019) Estimation of dissolved oxygen in riverine ecosystems: Comparison of differently optimized neural networks. Ecol Eng 138:298–309. https://doi.org/10.1016/j.ecoleng.2019.07.023
9. Dehghani R, Poudeh HT, Izadi Z (2021) Dissolved oxygen concentration predictions for running waters with using hybrid machine learning techniques. Model Earth Syst Environ. https://doi.org/10.1007/s40808-021-01253-x
10. Ghritlahre HK, Prasad RK (2018) Exergetic performance prediction of solar air heater using MLP, GRNN and RBF models of artificial neural network technique. J Environ Manage 223:566–575. https://doi.org/10.1016/j.jenvman.2018.06.033
11. Ghorbani MA, Zadeh HA, Isazadeh M, Terzi O (2016) A comparative study of artificial neural network (MLP, RBF) and support vector machine models for river flow prediction. Environ Earth Sci 75:476. https://doi.org/10.1007/s12665-015-5096-x
12. Heddam S, Kisi O (2017) Extreme learning machines: a new approach for modeling dissolved oxygen (DO) concentration with and without water quality variables as predictors. Environ Sci Pollut Res 24:16702–16724. https://doi.org/10.1007/s11356-017-9283-z
13. Heidari E, Sobati MA, Movahedirad S (2016) Accurate prediction of nanofluid viscosity using a multilayer perceptron artificial neural network (MLP-ANN). Chemom Intell Lab Syst 155:73–85. https://doi.org/10.1016/j.chemolab.2016.03.031
14. Hopfield J (1988) Artificial neural networks. IEEE Circuits Syst Mag 4(5):3–10. https://doi.org/10.1109/101.8118
15. Hossain R, Than AM, Ali S (2013) The Combined Effect of Applying Feature Selection and Parameter Optimization on Machine Learning Techniques for Solar Power Prediction. Am J Energy Res 1(1):7–16. https://doi.org/10.12691/ajer-1-1-2
16. Jordan MI, Mitchell TM (2015) Machine learning: Trends, perspectives, and prospects. Sci 349(6245):255–260. https://doi.org/10.1126/science.aaa8415
17. Keshtegar B, Heddam S (2018) Modeling daily dissolved oxygen concentration using modified response surface method and artificial neural network:a comparative study. Neural Comput Appl 30:2995–3006. https://doi.org/10.1007/s00521-017-2917-8
18. Kisi O, Alizamir M, Gorgij AD (2020) Dissolved oxygen prediction using a new ensemble method. Environ Sci Pollut Res 27:9589–9603. https://doi.org/10.1007/s11356-019-07574-w
19. Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97:273–324. https://doi.org/10.1016/S0004-3702(97)00043-X
20. Kumar ZM, Manjula R (2012) Regression model approach to predict missing values in the Excel sheet databases. Int J Comput Sci Eng Technol (IJCSET) 3(4):130–135
21. Li X, Sha J, Wang ZL (2017) A comparative study of multiple linear regression, artificial neural network and support vector machine for the prediction of dissolved oxygen. Hydrol Res 48(5):1214–

1225. https://doi.org/10.2166/nh.2016.149

22. Lord D, Qin X, Geedipally SR (2021) Highway Safety Analytics and Modeling, Exploratory Analysis of Safety Data, 1st Ed. Elsevier, Netherland. ISBN: 9780128168196
23. Ma Y, Ding W (2018) Design of Intelligent Monitoring System for Aquaculture Water Dissolved Oxygen. In IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC 2018), Chongqing, China. https://doi.org/10.1109/IAEAC.2018.8577649
24. Mahesh B (2020) Machine Learning Algorithms -A Review. Int J Sci Res (IJSR) 9(1). https://doi.org/10.21275/ART20203995
25. Maheshwari K, Lamba S (2019) Air Quality Prediction using Supervised Regression Model. In 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), Ghaziabad, India. https://doi.org/10.1109/ICICT46931.2019.8977694
26. Masmoudi S, Elghazel H, Taieb D, Yazar O, Kallel A (2020) A machine-learning framework for predicting multiple air pollutants' concentrations via multi-target regression and feature selection. Sci Total Environ 715:136991. https://doi.org/10.1016/j.scitotenv.2020.136991
27. Mehr HD, Polat H (2021) Diagnosis of polycystic ovary syndrome through different machine learning and feature selection techniques. Health and Technol 12:137–150. https://doi.org/10.1007/s12553-021-00613-y
28. Ouma YO, Okuku CO, Njau EN (2020) Use of Artificial Neural Networks and Multiple Linear Regression Model for the Prediction of Dissolved Oxygen in Rivers: Case Study of Hydrographic Basin of River Nyando, Kenya. Hindawi 2020(9570789):23. https://doi.org/10.1155/2020/9570789
29. Qiao W, Tian W, Tian Y, Yang Q, Wang Y, Zhang J (2019) The Forecasting of PM2.5 Using a Hybrid Model Based on Wavelet Transform and an Improved Deep Learning Algorithm. IEEE Access 7:142814–142825. https://doi.org/10.1109/ACCESS.2019.2944755
30. Rozemeijer JC, Broers HP (2007) The groundwater contribution to surface water contamination in a region with intensive agricultural land use (Noord-Brabant, The Netherlands). Environ Pollut 148(3):695–706. https://doi.org/10.1016/j.envpol.2007.01.028
31. Rustam F, Reshi AA, Mehmood A, Ullah S, On BW, Aslam W, Choi GS (2020) COVID-19 Future Forecasting Using Supervised Machine Learning Models. IEEE Access 8:101489–101499. https://doi.org/10.1109/ACCESS.2020.2997311
32. Shaghaghi N, Nguyen T, Patel J, Soriano A, Mayer J (2020) DOxy: Dissolved Oxygen Monitoring. In IEEE Global Humanitarian Technology Conference (GHTC), Seattle, WA, USA. https://doi.org/10.1109/GHTC46280.2020.9342916
33. Singh KP, 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
34. Valera M, Walter RK, Bailey BA, Castillo JE (2020) Machine Learning Based Predictions of Dissolved Oxygen in a Small Coastal Embayment. J Mar Sci Eng 8:1007. https://doi.org/10.3390/jmse8121007

35. Varol M, Şen B (2012) Assessment of nutrient and heavy metal contamination in surface water and sediments of the upper Tigris River, Turkey. CATENA 92:1–10. https://doi.org/10.1016/j.catena.2011.11.011
36. Wang Y, Li Y, Pu W, Wen K, Shugart YY, Xiong M, Jin L (2016) Random Bits Forest: a Strong Classifier/Regressor for Big Data. Sci Rep 6(30086):1–8. https://doi.org/10.1038/srep30086
37. Zebari RR, Abdulazeez AM, Zeebaree DQ, Zebari DA, Saeed JN (2020) A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. J Appl Sci Technol Trends 1(2):56–70. https://doi.org/10.38094/jastt1224
38. Zhu S, Heddam S (2020) Prediction of dissolved oxygen in urban rivers at the Three Gorges Reservoir, China: extreme learning machines (ELM) versus artificial neural network (ANN). Water Qual Res J 55(1):106–118. https://doi.org/10.2166/wqrj.2019.053

Figures

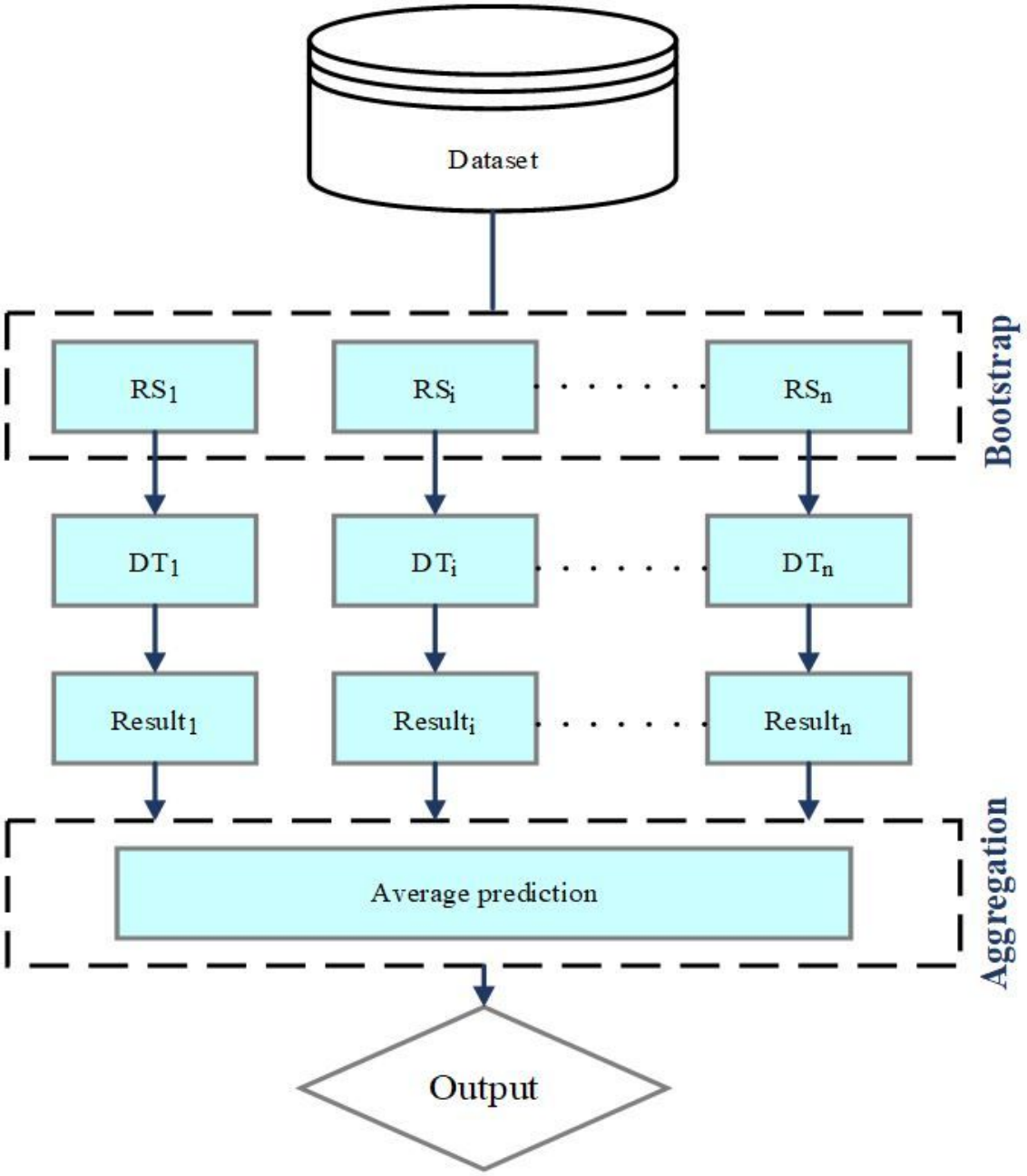

Figure 1

Schematic workflow of Random Forest (RF) regressor (RS: random sample; DT: decision tree).

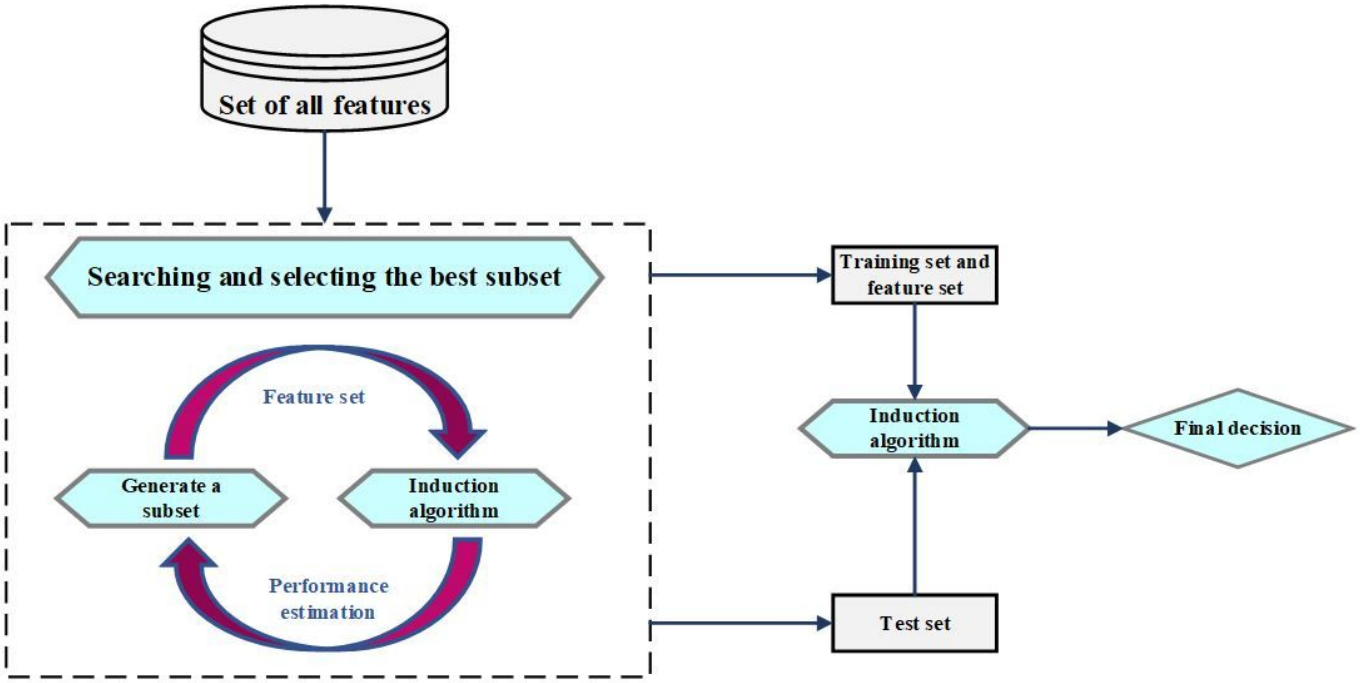

Figure 2

Schematic workflow of wrapper feature selection method.

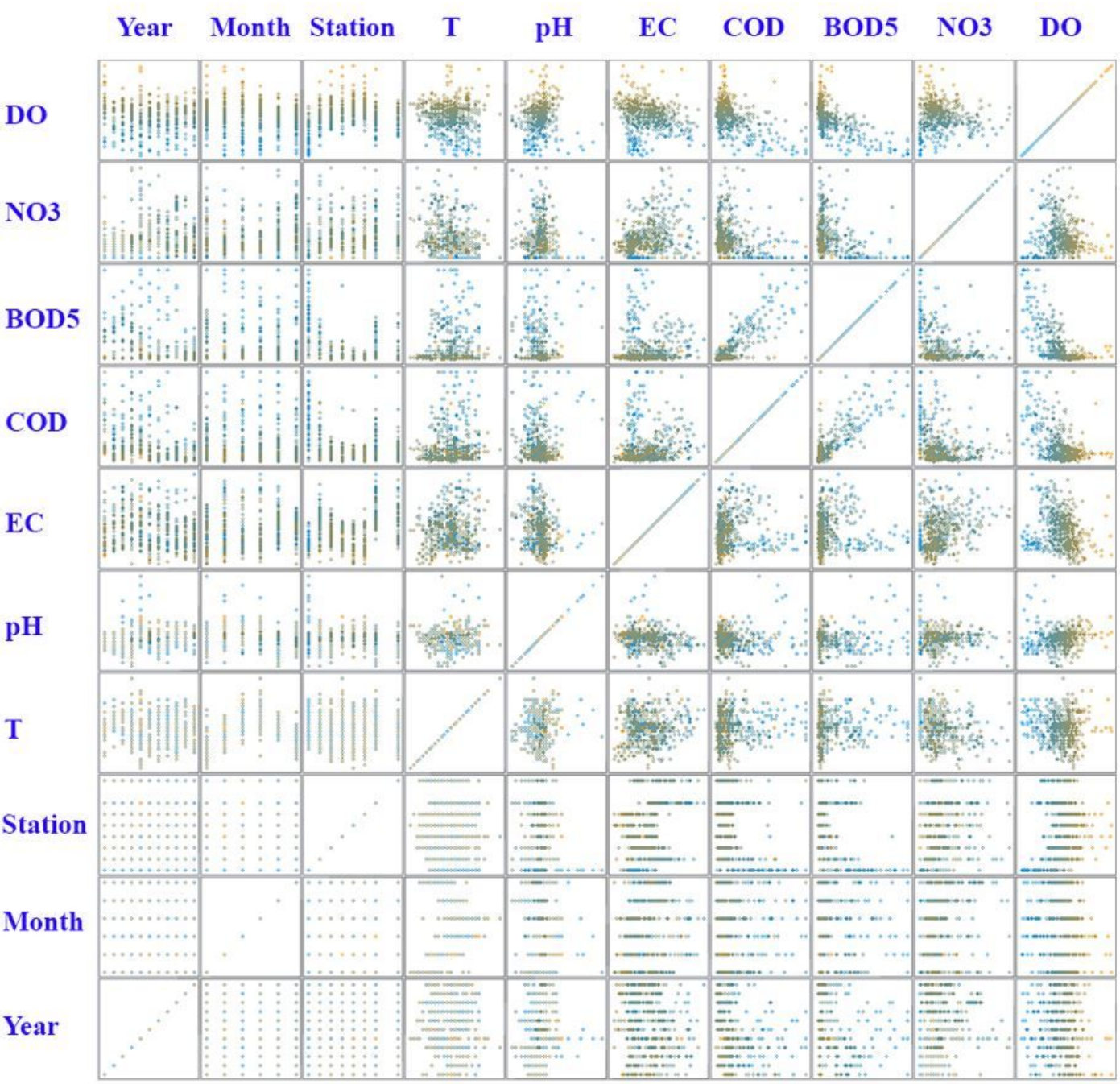

Figure 3

Correlation plot matrix of the variables selected by the wrapper feature selector.

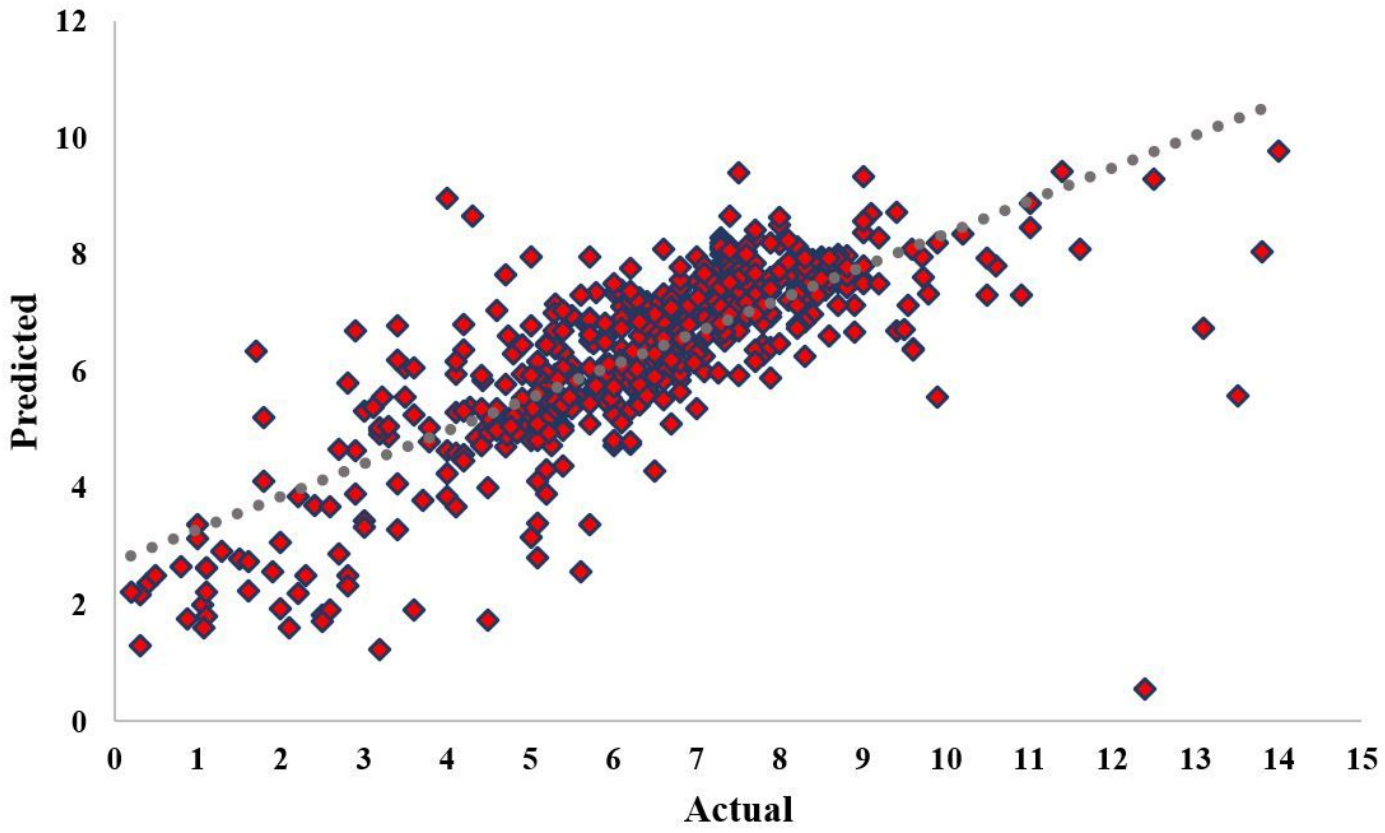

Figure 4

Predicted DO concentration by the RF regressor with the whole variables of interest against actual values of DO

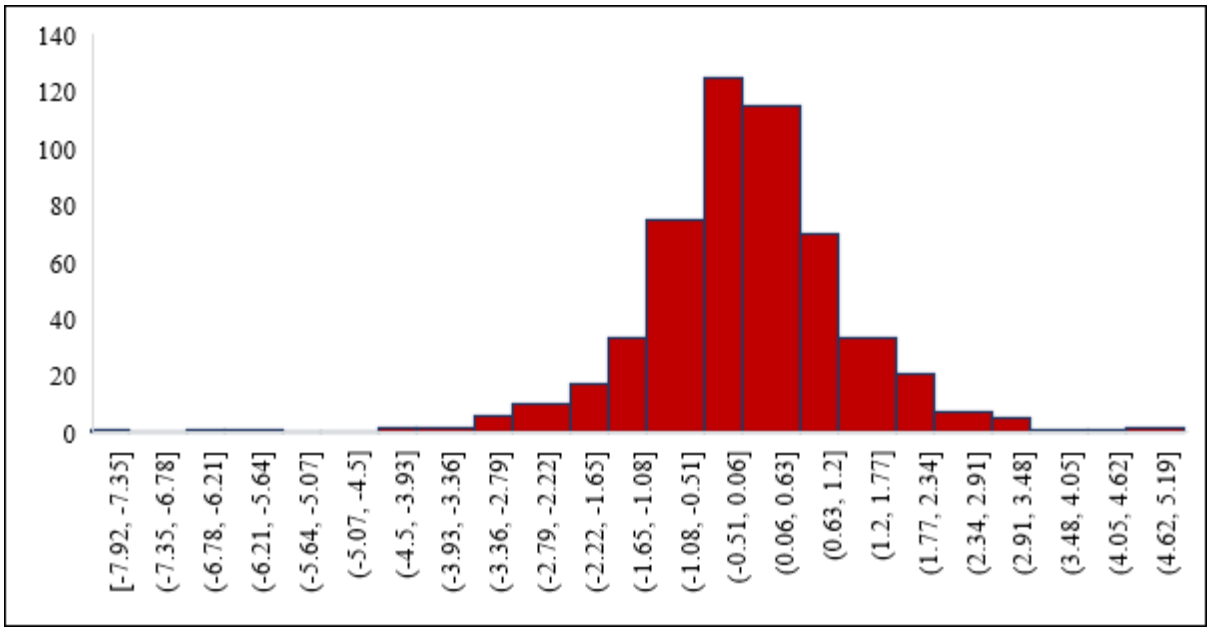

Figure 5

Testing phase estimation error deviation of the RF model trained by the whole variables of interest

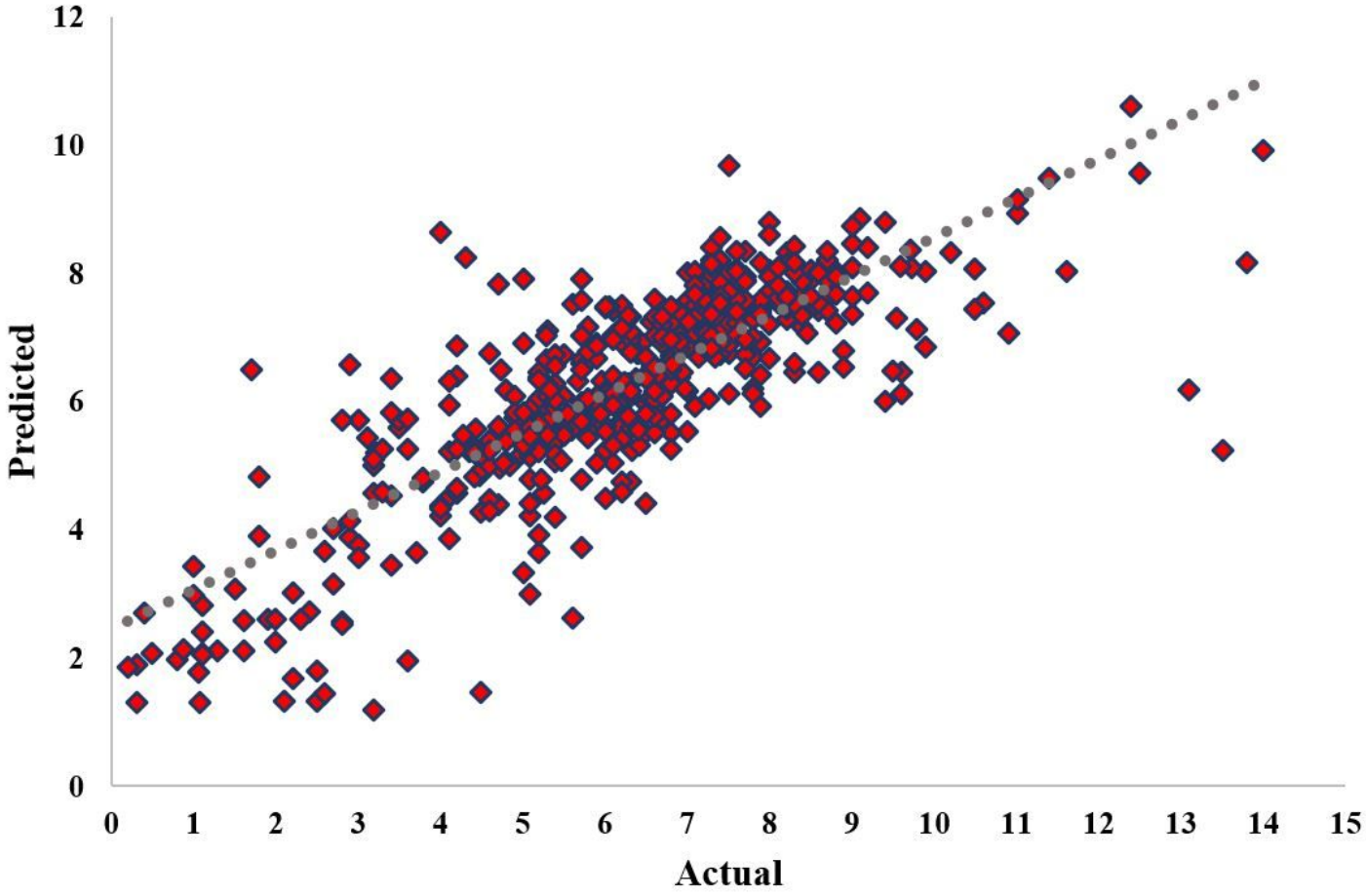

Figure 6

Predicted DO concentration by the RF regressor with the reduced subset of features against actual values of DO

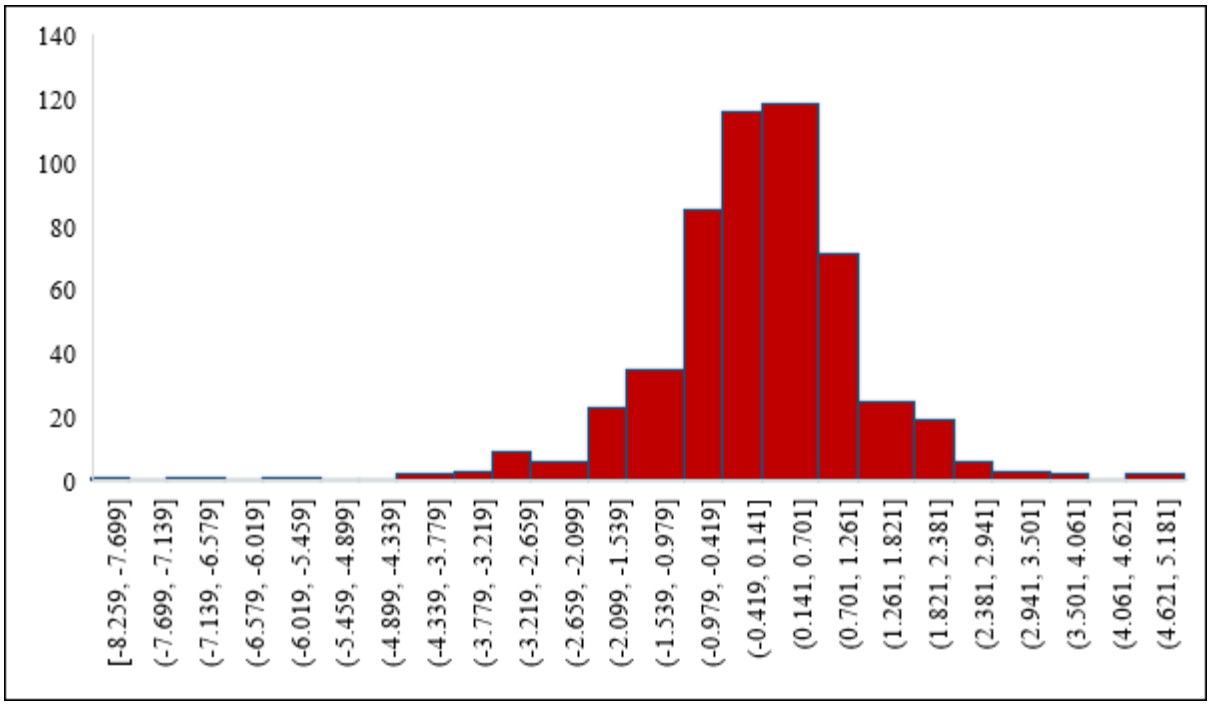

Figure 7

Testing phase estimation error deviation of the RF model trained by the reduced subset of features

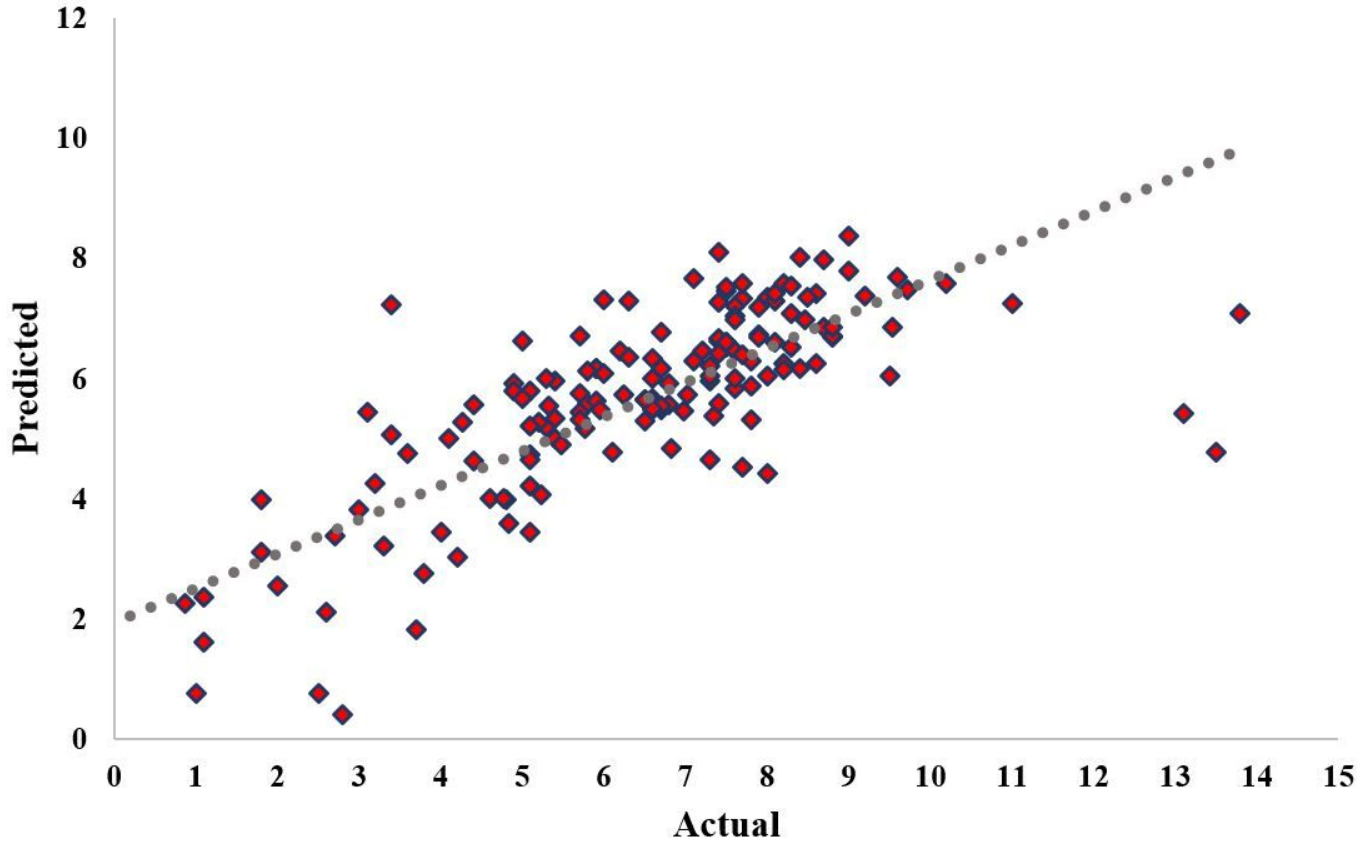

Figure 8

Predicted DO concentration by the MLP regressor with the whole variables of interest against actual values of DO.

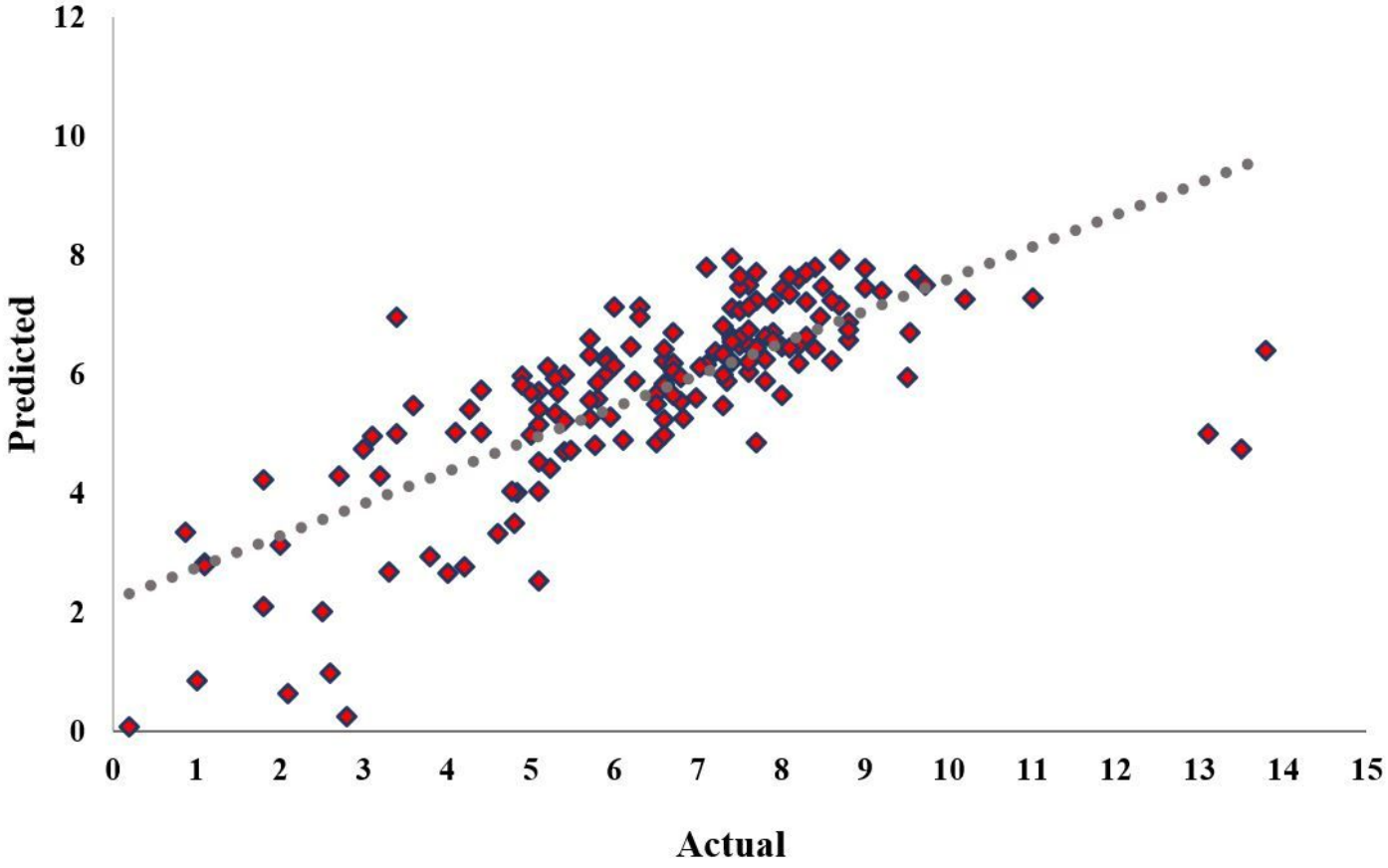

Figure 9

Predicted DO concentration by the MLP regressor with the reduced subset of features against actual values of DO.

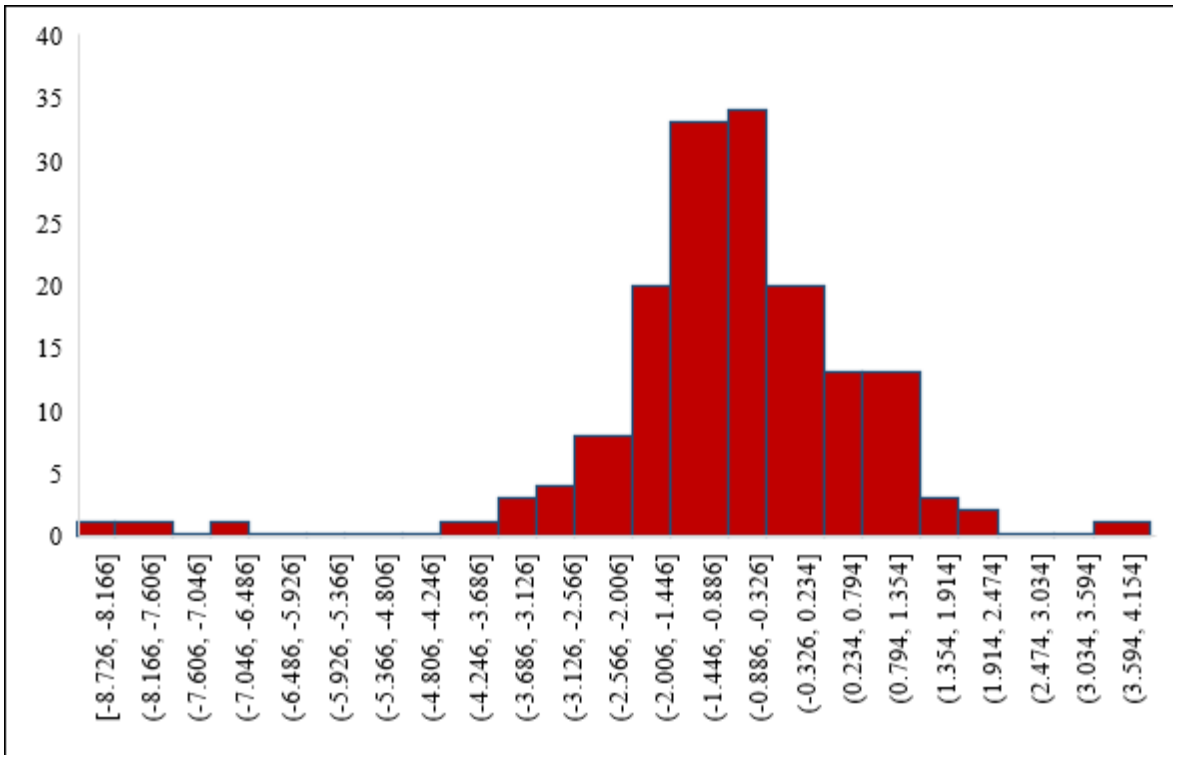

Figure 10

Testing phase estimation error deviation of the MLP model trained by the whole variables of interest.

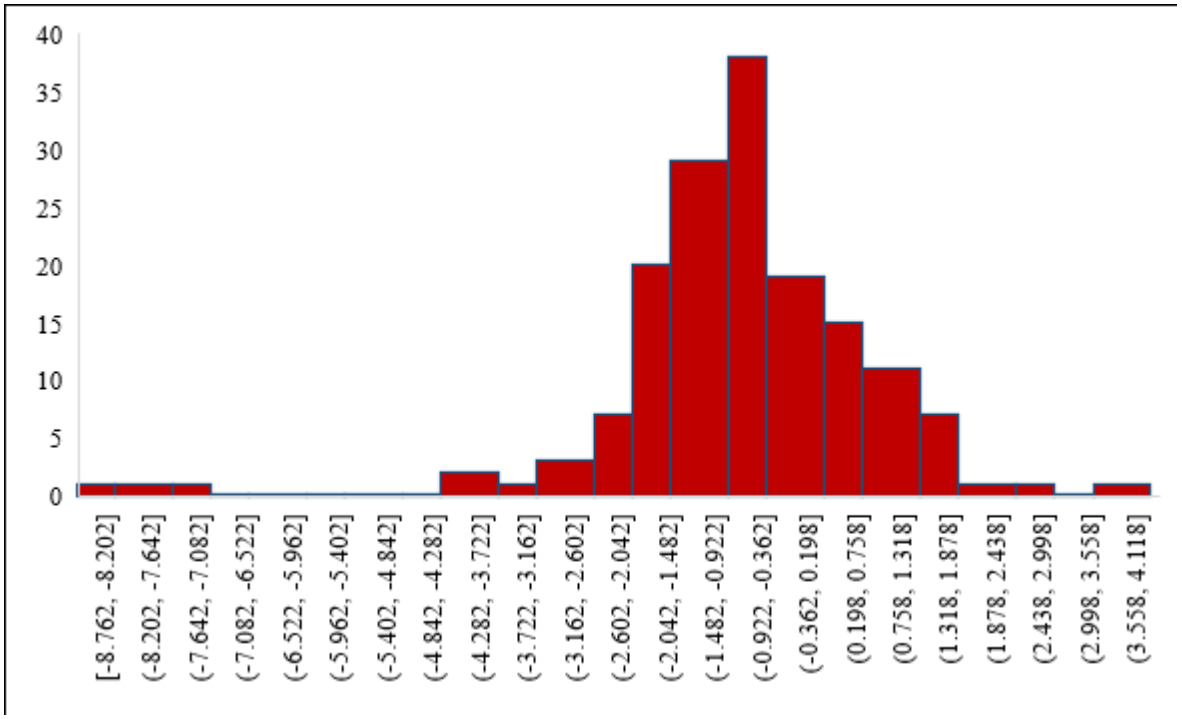

Figure 11

Testing phase estimation error deviation of the MLP model trained by the reduced subset of features.