Predictive Analysis of Water Quality Parameters using Deep Learning

Archana Solanki
Computer Science
Symbiosis Institute of Technology
Pune, 412115, India

Himanshu Agrawal
Computer Science
Symbiosis Institute of Technology
Pune, 412115, India

Kanchan Khare
Computer Science
Symbiosis Institute of Technology
Pune, 412115, India

ABSTRACT
Lakes and reservoirs are important water resources.
Reservoirs are vital water resources to support all living organism. They provide clean water and habitat for a complex variety of aquatic life. Water from such resources can be used for diverse purposes such as, industry usage, agriculture and supplies for drinking water and recreation and aesthetic value.

Apart from this, reservoirs also helpful to get hydro-electric power, flood control and scenic beauty. Water collected in such resources can be utilized in drought situation also. Unfortunately, these important resources are being polluted and the quality of water is being influenced by numerous factors. The quality of water is deteriorated by anthropogenic activities, indiscriminate disposal of sewage, human activities and also industry waste. Water quality monitoring of reservoirs is essential in exploitation of aquatic resources conservation. The quality of water helps in regulating the biotic diversity and biomass, energy and rate of succession. Moreover, contaminated water can lead to some waterborne diseases and also influences child mortality. In order to reduce effect of contaminated water, it is essential to assess different aspects of water quality. Predicting water quality parameters a few steps ahead can be beneficial to achieve this. The main objective of this study is to provide fairly accurate predictions for variable data. The research was carried out by using the secondary data collected from a third party for Chaskaman River located near Nasik, Maharashtra, India on WEKA tool. The study shows that deep learning techniques which use unsupervised learning to provide accurate results as compared to the techniques based on supervised learning. The comparison of results show that robustness can be achieve by denoising autoencoder and deep belief network and also successfully handle the variability in the data. Merit of the unsupervised learning algorithms are evaluated on the basis of metrics such as mean absolute error and mean square error to examine the error rate of prediction.

General Terms
Machine learning, Predictive analytics

Keywords
Deep Learning, Unsupervised learning, Deep belief network, Denoising auto-encoders, Restricted Boltzmann Machine

1. INTRODUCTION
Water is the most important resource for the entire humanity. Lakes and reservoirs are such water bodies whose judicious exploitation can directly affect the humanity as all the living organism depends on it. Hence, water bodies as a domain for planning and management has been accepted all over the world. Unfortunately, these useful resources are deteriorated by humans.

Anthropogenic activities, contamination of chemicals from industries, sediments etc. are the factors that affect the quality of the water directly. Due to increase in urbanization, the exploitation and deterioration of water bodies has been increased. Contaminated water resources can cause serious effects on human as well as aquatic life. Along with these, the quality of water is influenced by numerous factors like human activities, soils, geology etc. Moreover, the quality of water is deteriorated by indiscriminate disposal of sewage, human activities. Also, contaminated water can lead to some waterborne diseases and also influences child mortality. In order to reduce effect of contaminated water, it is essential to assess different aspects of water quality. Predicting water quality parameters a few steps ahead can be beneficial to achieve this. Hence, water quality monitoring of reservoirs is essential in exploitation of aquatic resources conservation. The quality of water helps in regulating the biotic diversity and biomass, energy and rate of succession.

Water pollution has been a major concern in many developing countries since decades. Contaminated water can lead to water scarcity which is another major issue [1]. There are various factors that may affect the quality of water. Such as, thermal pollution, acidification, salinization etc. Moreover, urbanization and industry effluents influence the surface water temperatures that affects the dissolved oxygen. Dissolved oxygen is vital to the metabolism living in the reservoir. pH of water can be influenced by effluents added by industries or human activities living around the water body. Whereas Salinization can reduce the amount of pure drinking water and can harm the aquatic life. Turbidity is an attribute of water which defines the purity of water. Higher transparency in water bodies shows the good quality of water. Reservoirs have high turbidity in monsoon when water is a bit murkier and also it gets affected by suspended solids. Temperature plays a vital role in water quality. It affects the chemical reaction speed and the aquatic metabolism. Aquatic metabolism has narrow temperature tolerance hence even moderate change in surface water temperature can harm the metabolism living in the water basin. Nutrients are essential elements to life. Bad quality of water can affect the nutrients and also the aquatic metabolism [2]. The data received by monitoring the quality of the water can be then useful to make predictions about the quality a few steps ahead in order to get better water management [3].

Recent report of United Nations reveals that the study can be helpful to limit the waterborne diseases and also it can help people to get pure and safe water for their daily needs. 3 million people in the world die of water related diseases due to contaminated water each year, including 1.2 million children. [4] Predictive analysis can help to capture relationships among many factors that can help to assess risk with a particular set of conditions. Predictive analysis includes data mining techniques, statistical analysis and modelling. In

this process, defining the objectives, deliverables and identifying the datasets is done at first. After defining the project, various data mining techniques are used to discover the useful information from the dataset. These extracted information is applied various data analysis and statistical analysis techniques. Finally, the predictive model is created in order to get the predictions. Multiple models are applied to the same dataset and the model which best fit is chosen [5].

Our contribution in this paper is as follows:

- Initial contribution of this dissertation comprises study of water quality challenges in reservoir using parameters such as pH, dissolved oxygen and turbidity.
- Predictive analysis of water quality on continuous water quality dataset from Chaskaman reservoir, Nasik using supervised learning such as ANN techniques.
- Further analysis using advance predictive technique, such as deep learning i.e. unsupervised learning.
- Performance analysis using metrics such as Mean squared error and Mean absolute error.

The paper has been divided into following sections; section two describes the related work done using different approaches. Section three contains the detail about the methods used in the research work followed by results and discussion and discussion.

2. LITERATURE REVIEW

Water quality is the priority criterion in matching water demand and supply. Natural and manmade lakes need to be monitored and managed to match the demand supply equation. A lot of initiative has been taken at national level [6], [7], [8].and underlines the use of water quality index as a quality indicator of water. The objective of developing an index is to simplify the complex water quality parametric data into comprehensive information for easy understanding. The latest review paper on water quality index includes various water quality indices (WQI) used in the surface water quality assessment [9].

At Sukhna Lake of Chandigarh city, the water quality index assessment has been initiated. The lake is a man-mane lake and is situated in the foothills of the Shivalik range of Chandigarh. The lake is utilized for various purposes like morning and evening strolls, sightseeing, boating and exercise. Hence, it is a primary tourist attraction in Chandigarh. The study of water quality index and its dependence on catchment characteristics has been carried out by using the National Sanitation Foundation Water Quality Index (NSFWQI) and Overall Index of Pollution (OIP) water quality classification. The water of the lake is contaminated by sewage and construction activities. In order to analyze the quality of water, the data were collected monthly from 2006 to 2012 manually. The deterioration of the quality of water, the administration of the city has announced strictly ban on new construction and also a sewage treatment plant has been started [10]. In 2014, some researchers of Malaysia conducted a study on two lakes named Chini and Bera. They collected samples from 2005 to 2009. The data sample consisted of 11 parameters which were used to predicate DO concentration. The DO concentration was dichotomized into three different levels such as, High, Medium and Low. They ranked the input parameters and they used forward selection method to determine the optimum parameters that yield the lowest errors and highest accuracy. The initial results showed that pH, temperature and conductivity significantly affect the prediction of DO. Then, they applied SVM model using the Anova kernel with those parameters yielded 74% accuracy rate. They concluded that using dichotomized value of DO yields higher prediction accuracy than using precise DO value and ANOVA is the most appropriate kernel to obtain the highest accuracy. [11]

At Terengganu River, Malaysia, a study was conducted to predict DO using SVM. They conducted the study for two different stations using the five parameters such as, pH, temperature, electrical conductivity and Nitrate and Ammonia Nitrogen. They used SVM with its non-linear and stochastic modelling proficiencies. The performance of the model was evaluated using three statistical indexed such as, Mean Squared Error (MSE), Coefficient of Efficiency (CE) and coefficient of Correlation (CC). They concluded that SVM can give robust and precise result and able to give fairly accurate predictions. It can also help in optimizing the water quality monitoring programs. [12] Recently in Malaysia, at a free constructed wetland, a research study was conducted to predict the WQI. They monitored seventeen points of the wetland were monitored. The sampled were monitored twice a month over a period of 14 months, and an extensive dataset was collected for 11 water quality variables. They have showed the detailed comparison of overall performance of AVM with coefficient of correlation (R2) = 0.9984 and Mean absolute error (MAE) = 0.0052 was either better or comparable with neural networks they concluded that these methods simplify the calculation of the WQI and reduce substantial efforts and time by optimizing the computations. [13] The model for predicting the WQI using feed-forward three layer perceptron has been develop at Kinta River (Malaysia). The ANN model was developed using multi-layer perceptron (MLP) consisting of two major steps, determining the network architecture and specifying network structure. The network was constructed using three layers, an input layer, one hidden layer and an output layer. The WQI predictions, generated using this model show that ANN approach in this particular field is more effective and satisfying. And therefore in encourages the use of ANN based approach in this area as comprehensive and highly reliable technique [14]. A comparative study of ANN algorithms was carried out to predict water quality for the river Ganga in 2014. The research was mainly emphasized on the supervised learning techniques of ANN. They used MATLAB neural network toolbox to create the neural network. The network was developed by MLP back propagation algorithm integrated with Levenberg Marquaradt (LM). Along with it, they used Gradient descent adaptive algorithm for prediction. The tangent hyperbolic function was used as an activation function for hidden layer whereas linear activation function was used at the output layer as an activation function. The findings show that LM performs better than GDA algorithm. It encourages to use ANN model as it is cost-effective and easy to use. [15]

In 2014, some researchers in Pakistan conducted a study on managing the water quality of watersheds. The study was conducted at Rawal watershed in Pakistan. It is relatively small water resource and being affected by anthropogenic activities like urbanization or deforestation. They collected monthly data from the watershed in order to analyze the quality of the water as per WHO standards. Regression model was applied to analyze he seasonal trends in water quality whereas the combination of supervised and unsupervised machine learning techniques were applied to test the quality indices. The study found that unsupervised learning

techniques like average linkage method of hierarchical clustering using Euclidean distance is an accurate method to find out the water quality index. Similarly, for classifications, they found MLP, a supervised learning technique can give accurate results. [3]

Table 1: Table of Notations

ANN	Artificial Neural Network
SVM	Support Vector Machine
WQI	Water Quality Index
MLP	Multi-layer Perceptron
SdA	Stacked Denoising Auto-encoder
DBN	Deep Belief Network
DO	Dissolved Oxygen
Wi	Weight of inputs
Si	Weighted sum of inputs
Xj	Inputs

3. METHODOLOGY

Methodologies adopted to get predictions for water quality has been discussed in this section. Deep learning based predictive analysis techniques were used.

- Data collection
  Data collection is the process of collecting data and discovering the patterns and behavior of the data. In order to achieve this, the secondary data for various parameters of water was collected from a third party. The data includes the parameters like temperature, pH, dissolved oxygen, turbidity, chloride etc. The data was collected from Krishna river basin near Chaskaman. The research was conducted by extracting three parameters pH, dissolved oxygen and turbidity from the dataset. These three parameters are the impactful attributes of water quality as they directly affect the living organism and hence, the quality of the water can be measured effectively using them.
- Data Pre-processing
  Data pre-processing helps to improve the quality and efficiency of the data. As raw data is inconsistent and noisy which affects the quality of the data. Data preprocessing is a step which deals with data preparation and transformation.
- Modelling
  This includes the process of developing a predictive model, testing and validating it. Deep learning based techniques were used in order to develop the model for prediction. In this research, denoising auto-encoder and deep belief network methods were used to develop the model. Both of these methods use unsupervised learning to train the model.

3.1 Stacked Denoising Autoencoder

The Stacked Denoising Auto-encoder can be formed by stacking denoising auto-encoders. It is an extension of the stacked auto-encoder. The unsupervised pre-training in SdA is done one layer at a time in which each layer is trained as a denoising auto-encoder by minimizing the reconstruction error. Once all layers are pre-trained, the network goes through a second stage of training called fine-tuning. The supervised fine-tuning is done to minimize prediction error on a supervised task. In order to achieve this, first the logistic regression layer is added on the top of the network. Then the entire network is trained as the training done in MLP. The stacked autoencoder has two facades: a list of auto-encoders and an MLP. The first facade is used during the pre-training whereas the second facade is used during the second training process. These two facades are linked because,

- The auto-encoder and sigmoid layers of the MLP share parameters
- The latent representations computed by intermediate layers of the MLP are fed as input to the autoencoders

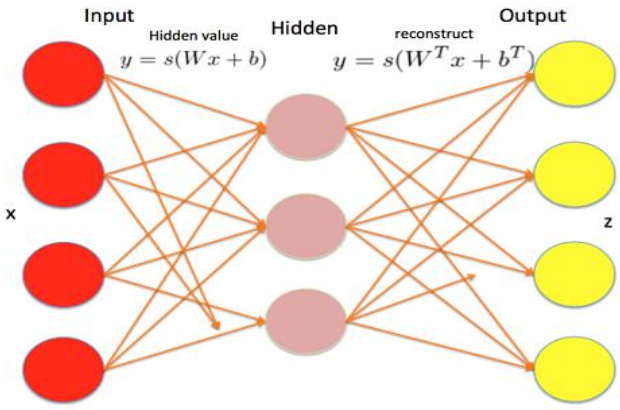

Fig 1: Denoising auto-encoder

3.2 Deep Belief Network

Deep belief network are the deep neural networks with many hidden layers which are connected with each other but the units of the layers are not connected. DBN are constructed by using restricted Boltzmann machine arranged in a stack. [14] Deep belief network model the joint distribution between the visible units x and l hidden layers h_k by using,

$$P(x, h^1, \dots, h^l) = \left(\prod_{k=0}^{l-2} P(h^k | h^{k+1})\right) P(h^{l-1}, h^l)$$

Where $x = h^0$; $P(h^{k-1}|h^k)$is a conditional distribution for the visible units.

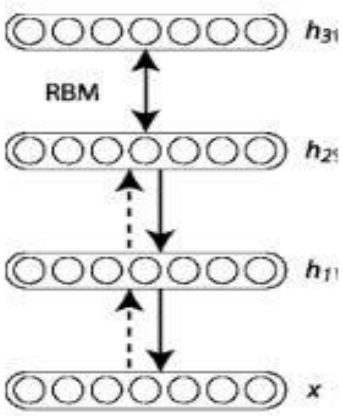

Fig. 2: Deep belief network

The principle of greedy layer-wise unsupervised training can be applied to DBNs with RBMs as the building blocks for each layer [23], [24]. The process is as follows:

Greedy layer wise unsupervised pre-training helps to overcome the challenges of deep learning prior to supervised fine-tuning. [16] Fine-tuning is then performed via supervised gradient descent of the negative log-likelihood cost function. [17] To develop the predictive model, continuous deep belief network was used. Continuous deep belief network is an extension of deep belief network. It uses continuum of decimals instead of binary data. [18] After training the test data, feed forward neural network was applied to get the prediction. The algorithm was developed using java. To compare the results of deep learning algorithm based predictions, some classical methods like MLP and linear regression were used.

3.3 Linear Regression

Linear regression is a classical statistical method

Algorithm 1 Greedy layer-wise unsupervised pre-training

STEP 1: Train the first layer as an RBM that models the raw input x = h(0) as its visible layer.

STEP 2: Use that first layer to obtain a representation of the input that will be used as data for the second layer. Two common solutions exist. This representation can be chosen as being the mean activations p(h(1) = 1jh(0)) or samples of p(h(1)jh(0)).

STEP 3: Train the second layer as an RBM, taking the transformed data (samples or mean activations) as training examples (for the visible layer of that RBM).

STEP 4: Iterate STEP 2 and 3 for the desired number of layers, each time propagating upward either samples or mean values.

STEP 5: Fine-tune all the parameters of this deep architecture with respect to a proxy for the DBN log- likelihood, or with respect to a supervised training criterion (after adding extra learning machinery to convert the learned representation into supervised predictions, e.g. a linear classifier).

which comes under regression analysis. Regression analysis is a technique which fits the straight line to patterns of data. In this method, he weights are calculated from training data [19]. Linear regression is supervised leaning based approach in which the dependent variable is predicted from the other independent variables [20]. Linear regression is the regression analysis method. It used linear combination of inputs. Linear regression uses a function

$$f(x) = w_0 + \sum_{j=1}^{d} w_j x_j$$

Linear regression calculates weights from the training dataset. To minimize the squared error, it uses

$$\sum_{i=1}^{n} \left(x^i - \sum_{j=0}^{k} w_j a_j^i \right)^2$$

Linear regression is well found venerable regression analysis method [19]. Such model assumes that the random variables are nonlinear and there is no autocorrelation between them. The regression line of the predicted value of Y from X passes through the origin and it has a slope which is equal to the correlation of X and Y [20].

3.4 Multi-Layer Perceptron

Single layer neural networks can handle only linear data. Hence, to overcome this limitation, a neural network with multiple layers is used [21]. Perceptron is the simplest form of neural networks. It consists of an input layer, hidden layer(s) and an output layer. Every neuron is connected and having a weight. It uses back-propagation to train the dataset [22]. MLP uses supervised learning approach. The function, applied over weighted sum

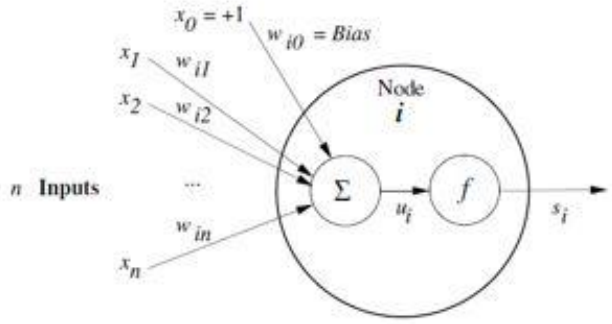

Fig. 3: Activation Function

of inputs is,

$$S_i = \left(W_{i,0} + \sum_{j \in l} W_{i,j} * S_j \right)$$

Error is calculated using [23] $e = d - d$, Where d = *desired value* and d= *estimated value*:

Mean Absolute Deviation (MAD) = $\frac{\sum_{i=1}^{N} |e_i|}{N}$

Sum Squared Error (SSE) = $\sum_{i=1}^{N} e_i^2$

Mean Squared Error (MSE) = SSE/ N

Root Mean Squared Error (RMSE) = $\sqrt{MSE}$

MLP is the most used type of nonlinear neural network. It can adapt changes according to the environment and maintains robustness. In MLP, Backpropagation algorithm is used as training algorithm. It uses one hidden layer for nonlinear dataset. It is used mainly to overcome over-fitting and under-fitting limitation of regressive analysis

4. EXPERIMENTS

The methods used in this paper are based on supervised and unsupervised learning approach. A secondary data was collected from a third party in which the parameter turbidity has more variation compared to pH and dissolved oxygen. After collecting the data, clustering technique was applied on it. Three clusters were created based on seasons; Winter, Summer and

Monsoon. Then data cleaning was applied, in which missing values were replaced by using the mean of available values. After applying data mining, results of classical methods were gathered using Weka tool and the results of deep learning algorithms were collected by using the code developed in java. The findings collected by the research are discussed in the next section.

5. RESULTS AND DISCUSSION

The graphs show that the method proposed in this paper work better for the data with high variability. Turbidity is the parameter which has high variability among all the three parameters. Results show that deep learning approaches are able to handle the under and over fitting of predictions as compared to the classical approaches such as MLP and Linear regression.

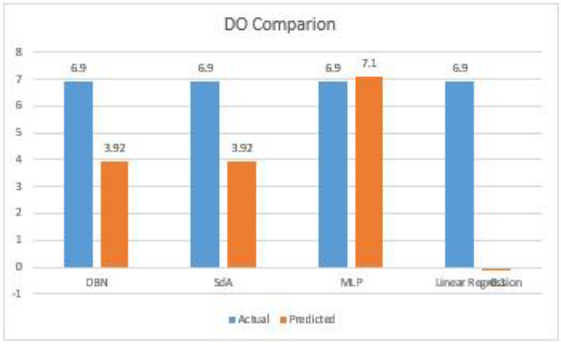

Fig. 4: Graph of Dissolved Oxygen

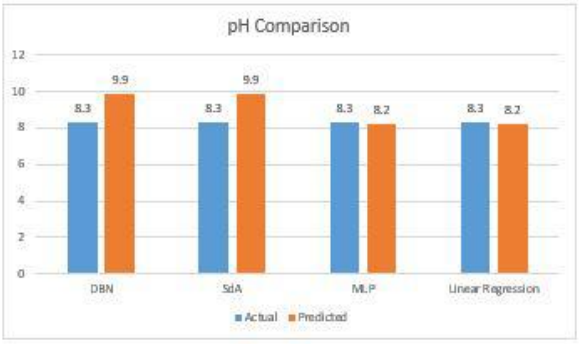

Fig. 5: Graph of pH

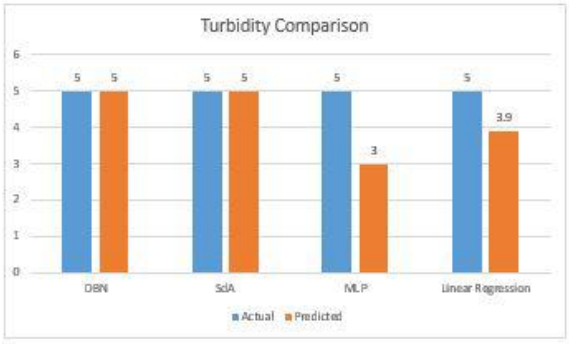

Fig. 6: Graph of Turbidity

6. CONCLUSION

The major conclusions derived from the water quality prediction modelling using deep learning approach are outlined below.

1) The results carried out from the study concludes that using unsupervised learning, data with variation can be predicted at acceptable accuracy rate

2) Results show that turbidity has high variation compared to the other two parameters but now is low. It is affected during the monsoon season most

3) pH has not much variation in data and hence, it is stable as compared to turbidity and DO

4) DO has a little variation during summer as the temperature affects the water quality during summer

This system can be implemented on system to continuously monitor the quality of the water. It can be helpful to monitor the quality of water in any uncertain condition.

The performance comparison of the algorithms has been shown in the following table:

7. REFERENCES

[1] D. Kriplani. etal., "Keeping the basin full Smart water for the 21st century," TCS.

[2] G. M. Carr. etal.,, "Water Quality for Ecosystem and Human Health,"UNESCO, Ontario, Canada, 2008.Fröhlich, B. and Plate, J. 2000.

[3] M. Ali. etal.,, "Data Analysis, Quality Indexing and Prediction of Water Quality for the Management of Rawal Watershed in Pakistan," in Islamabad, Pakistan, 2008.

[4] Poor water quality, a serious threat, [Online]. Available at: http://www.deccanherald.com/content/63740/poorwater-quality-serious-threat.html.

[5] Predictive analysis World, [Online]. Available at: http://predictiveanalyticsworld.com/predictiveanalytics.php.

[6] Y. Papadimitris. etal.,, "Integrated approach of lake quality monitoring", 2005.

[7] "Planning of water quality monitoring systems Technical document UNEP/GEMS Water", 2012.

[8] "Status of water quality in India -2010, Central pollution control board, Ministry of Environment and forests", 2010.

[9] P. Tirkey. etal.,, "Water quality indices- important tools for water quality assessment,"IJAC, Vol. 1, pp. 15-28, 2013.

[10] P. Chaudhry. etal.,, "Water Quality Assessment of Sukhna Lake of Chandigarh City of India; Hydro Nepal," Vol. 12, pp. 26-31, 2013.

[11] S. Malek. etal.,, "Dissolved Oxygen Prediction Using Support Vector Machine," Vol. 8, pp.153-160, 2014.

[12] A. Tarmizi. etal.,, "Dissolved Oxygen Prediction Using Support Vector Machine in Terengganu River," pp.2182-2188, 2014.

[13] R. Mohammadpour. etal.,, "Prediction of water quality index in constructed wetlands using support vector machine," Vol. 22, pp.6208–6219, 2015.

[14] H. Juahir. etal., Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors, Marine Pollution Bulletin, (2012).

[15] A. Giri. etal., Comparison of Artificial Neural network algorithm for water quality prediction of River Ganga, Environmental Research journal 8(2):55-63 (2014).

[16] D. Erhan. etal., Why Does Unsupervised Pre-training Help Deep Learning?, Journal of Machine Learning research 11: 625-660 (2010).

[17] Deep Learning tutorial, [Online]. Available at: http:/deeplearning.net.watersheds, [Online].

[18] Availableat:http://water.epa.gov/type/watersheds/monitoring/monintr.cfm.

[19] I. Witten [Class 4 – Lesson 1 Classification boundaries], University of Waikato, New Zealand).

[20] Nau, "Notes on linear regression analysis", [Online]. Available at: http://people.duke.edu/rnau/notesonlinearregressionanalysis.pdf.

[21] Piedmiller, "Machine Learning: MultiLayer Perceptrons", [Online]. Available at: http://ml.informatik.unifreiburg.de/media=teaching=ss10/05mlps.printer.pdf.

[22] I. Witten [Class 5 – Lesson 1 Simple neural networks], University of Waikato, New Zealand).

[23] G. Hinton. etal., "Reducing the Dimensionality of Data with Neural Networks,' ' 313(5786): 504 - 507 (2006).

[24] P. Bengio. etal., "Greedy Layer-Wise Training of Deep Networks, in Advances in Neural Information Processing Systems 19 (NIPS'06),' '153-160 (2007).

[25] Paulo Cortez, "Multilayer Perceptron (MLP) Application Guidelines", [Online]. Available at: http://www.dsi.uminho.pt/pcortez.F,gmnf,gmndf,mgnfd,gndf,mng.

Method	**Attribute**	**Predicted Value**	**Actual Value**	**MAE**	**MSE**	**SE**
LR	Turbidity	3.9	5	1.7	3.9	-
	pH	8.2	8.3	0.1	0.01	-
	DO	-0.1	6.9	0.3	0.16	-
MLP	Turbidity	3	5	0.02	0.01	-
	pH	8.2	8.3	0.0001	0	-
	DO	7.1	6.9	0.001	0	-
sDA	Turbidity	5	5	-	-	0
	pH	9.9	8.3	-	-	-1.6
	DO	3.92	6.9	-	-	2.98
DBN	Turbidity	5	5	-	-	0
	pH	9.9	8.3	-	-	-1.6
	DO	3.92	6.9	-	-	2.98

Table 2: Comparative results of the experiments