Using Machine Learning Approaches to Predict Water Quality of Ibn Battuta Dam (Tangier, Morocco)

El Mustapha Azzirgue[1], Farida Salmoun[1], El Khalil Cherif[2], Taha Ait Tchakoucht[3] and Nezha Mejjad[4]

[1]*Physico-chemistry Laboratory of Materials, Natural Substances and Environment (LAMSE), Faculty of Sciences and Techniques of Tangier, Morocco*

[2]*Institute for Systems and Robotics, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal*

[3]*School of Digital Engineering and Artificial Intelligence, Euromed University, Fes, Morocco*

[4]*Department of Geology, LGAGE, Faculty of Sciences Ben M'Sik, Hassan II University, B.P 7955, Casablanca 20670, Morocco*

Keywords: Water Quality, Dissolved Oxygen, Machine Learning, Ibn Batouta Dam.

Abstract: Ibn Batouta dam was built in 1977 next to catchment outlet and provides the Tangier-Assilah cities inhabitants with drinking water. The present study aims to assess the quality of Ibn Batouta dam water and use machine learning approaches to predict the future water quality of this dam. We select Dissolved Oxygen as the vital quality factor to be forecasted. A Long-short term memory (LSTM) network and a fully connected multilayer perceptron (MLP) are employed to predict Dissolved Oxygen in the next five years. The two models are assessed concerning, for data collected over twenty-one years (between 1998 and 2019) from Ibn Battuta station. The two models performances are compared and evaluated based on three metrics, Mean Square Error (MSE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Experimental results show that LSTM outperforms MLP by reducing RMSE, MSE and MAE respectively by 87%, 98% and 88%, indicating that the LSTM model is more accurate in tackling time series.

1 INTRODUCTION

Water makes up about 70% of the globe's surface and is vital for sustaining life (Umair Ahmed et al., 2019). Dam water is an essential source for the supply of drinking water, irrigation, etc. Besides their essential roles in agriculture, dam water offers special conditions for aquatic flora and fauna (Ouhmidou et al., 2015). In Morocco, the growing water needs for irrigation, electricity production, and the supply of drinking water have necessitated constructing many dams (Mohamed ACHMIT et al. ,2017). Due to urbanization expansion, population growth and intensification of agriculture, freshwater has become the most endangered ecosystem type in large parts of the world, experiencing widespread historical and continuing declines in land use quantity and quality of habitats and abundance of many species (S. Jannicke Moe et al. ,2019).

The use of conventional methods to assess the quality of surface water is generally expensive and time consuming. However, applying the artificial intelligence models can overcome this problem by predicting and evaluating water quality using Physico-chemical parameters as characteristics (Ali El Bilali et al. ,2021). Artificial intelligence algorithms for modelling water quality have been explored in recent years (Amir Mosavi et al., 2018). These algorithms explore the hidden and complex relationships between input and output variables to generate models that best represent these relationships. Numerous advantages of Artificial intelligence models over traditional physical and statistical models are as follow: the data needed for Artificial intelligence models can be collected relatively in an easy way, sometimes from remote sensing platforms; Artificial intelligence models are less sensitive compared to traditional approaches to missing data; the structures of Artificial intelligence models are flexible, non-linear and robust; and Artificial intelligence models can process enormous amounts of data at different scales (Costabile, P. et

Azzirgue, E., Salmoun, F., Cherif, E., Ait Tchakoucht, T. and Mejjad, N.
Using Machine Learning Approaches to Predict Water Quality of Ibn Battuta Dam (Tangier, Morocco).
DOI: 10.5220/0010734100003101
In *Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning (BML 2021)*, pages 350-354
ISBN: 978-989-758-559-3

al. 2013) (Fernández-Pato, J et al. 2016). This work aims at forecasting Dissolved O2 for up to the next five years, based on historical values of (pH, T °, Conductivity, Dissolved Oxygen, MES, PT, Chl a, NO2-, NO3-, NH4+, PO43-, SO42-) parameters, hence predicting the quality of the dam water Ibn Batouta, using artificial intelligence. The results obtained are positively promising, and the proposed approach offers an effective alternative to calculate and predict the water quality of the Ibn Batouta dam.

2 PRESENTATION OF THE STUDY AREA

Our study area, located at the South-East of Tangier (far from Tangier, about 17 km) (Azzirgue EM, and Salmoun F, 2019). The commune of Sebt Zinat houses the Ibn Battuta dam, a source of drinking water for Tangier. The town belongs to the Tanger Assilah province, part of the Tanger-Tetouan Al Hoceima region. It is bounded to the north by el Aouama commune, to the west by el Menzela commune, to the east by Jouamaa commune and to the south by Dar Chaoui commune (SETRAGEC 2018).

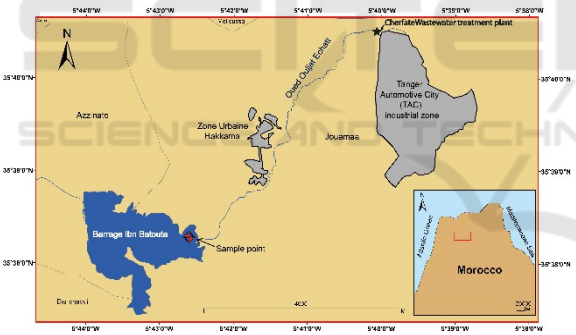

Figure 1: Geographical location of the study area and samples points.

3 PROPOSED APPROACH

3.1 Multi-Layer Perceptron (MLP)

The multilayer perceptron is the most famous feed-forward artificial neuron network (signals are transferred in one direction). It mainly comprises three components, an input layer containing input features, a hidden layer of neurons, and an output layer. Neurons are small computing units that use a non-linear function called the activation function. MLP adopts a supervised learning technique (output is known) that uses backpropagation to train the Model. MLP is different from a simple perceptron in the way that it can separate data that are not linearly separable and thus is appropriate for complex classification and regression problems. (Marius et al., 2009).

3.2 Long-Short Term Memory (LSTM)

Long-Short Term Memory artificial neural network is a category of recurrent neural networks that is a well-suited architecture for dealing with sequence data, as it has memory (feedback) connections and hence can catch temporal dependencies. LSTM was developed to overcome the problems of vanishing gradient encountered with traditional RNN models and the issue of short-time memory. LSTM is mainly composed of 4 components, a cell that memorizes sequences of previous time steps as well as three additional gates, namely, input, output, and forgets gates that manage information transmission to and from the cell (Hochreiter and Schmidhuber 1997).

3.3 The Architecture of the Approach

Dissolved Oxygen forecasting architecture is built upon the following steps, as shown in figure 2.

Step1: Data pre-processing. Since data is time series, it should be reframed to a supervised learning problem to predict Dissolved oxygen at the current year (t), given Dissolved oxygen and other parameters at the initial time step (t-1).

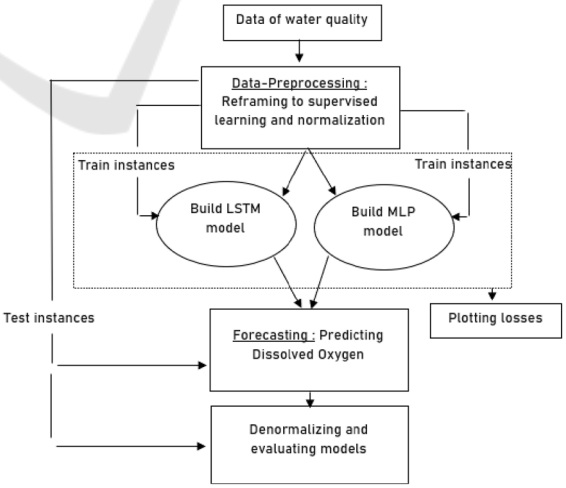

Figure 2: Architecture of the approach.

For this purpose, the models will be trained on data from 15 years and applied to forecast the five remaining years in the reframed dataset, given the

fundamental values of Dissolved Oxygen. Data is normalized to speed up the convergence of gradient descent algorithm using Min-Max technique applied on each input parameter according to the equation (1):

$$x = \frac{x - min}{max - min} \quad (1)$$

Step2: Splitting normalized data into training and testing sets and feeding them to two separate models, namely LSTM and MLP, then train the models for a several epochs, with weights initialized randomly. Train and test losses are then plotted.

Step3: The trained models are used to forecast Dissolved Oxygen up to five years in the future (number of observations in the test set). Predicted values are combined with input values, and then normalization is inverted to get back to the original scaling. The same is done for actual values combined with input values. Predicted and actual Dissolved oxygen values in their initial scaling are extracted, and performance metrics (rmse, mse, mae) are evaluated.

4 EXPERIMENTS AND RESULTS

4.1 Experimental Data

Data used in the analysis were collected over twenty one years (between 1998 and 2019) from Ibn Battuta station (21 observations, one observation for each year). Each instance in the dataset refers to a specific year, in which water quality is measured in a delimited area in the dam, based on 12 parameters (pH, T °, Conductivity, Dissolved O2, MES, PT, Chl a, NO2--, NO3-, NH4 +, PO43-, SO42-). Experiments were performed using Keras, and sci-kit-learn frameworks, in a PC 64-bit OS with 8GB RAM and Intel i5 1.6GHz, 1.8GHz. Figure 3 shows time series plots for each input parameter. Dissolved Oxygen is selected as the critical factor to be forecasted.

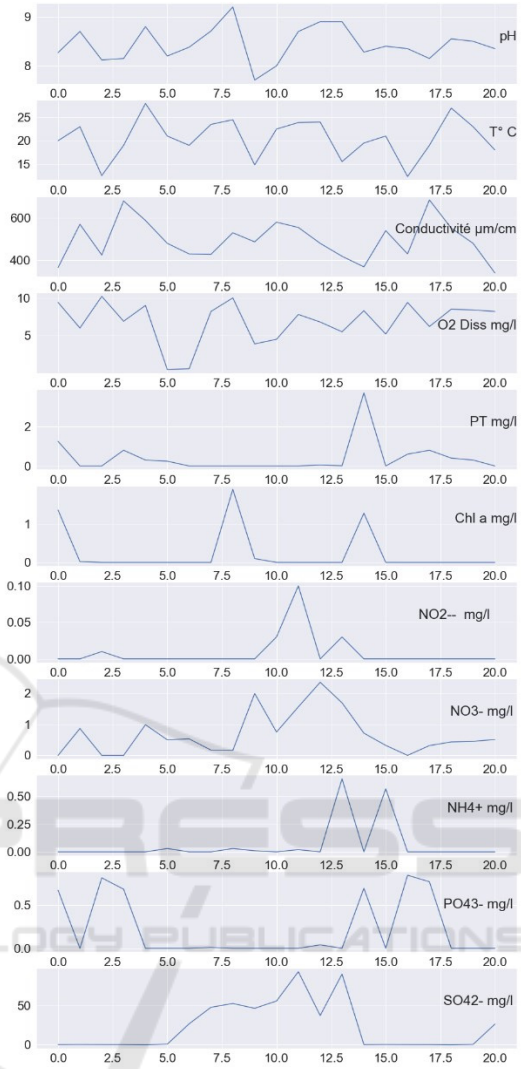

Figure 3: Time series of involved parameters.

4.2 Results and Discussion

4.2.1 Qualitative Results

LSTM and MLP Learning techniques are used to build two models to perform Dissolved Oxygen prediction and then are compared to adopt the most accurate approach. The best results were found for the following network configurations. We set MLP network with one hidden dense layer of 100 neurons and one dense output layer for predicting Dissolved Oxygen. Weights are randomly initialized. Stochastic gradient descent (sgd) is used as an optimizer. The training was performed with 30 epochs and tested over a test set at each period. Figure 4 show train and test losses for the MLP model, with network as mentioned above.

We forecast Dissolved Oxygen for the next five years given prior Dissolved Oxygen and other input parameters values. Figure 5 shows the time series for both actual values (ground truth) and predictions for the next five years (the 1st year is 0 in the x-axis).

LSTM was configured to include one hidden layer with 100 units and one dense output unit for prediction. Training is performed with 100 epochs and a batch size of 5 samples and a learning rate of 0.001. Weights are randomly initialized.

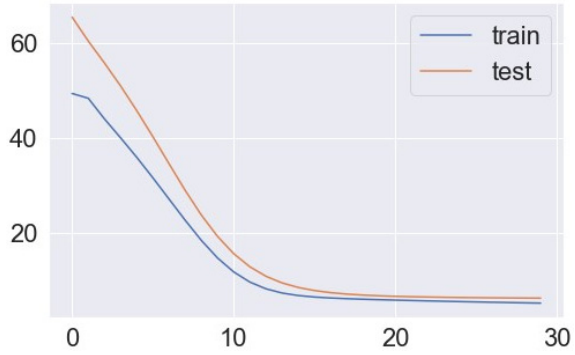

Figure 4: Train and Test losses for LSTM model.

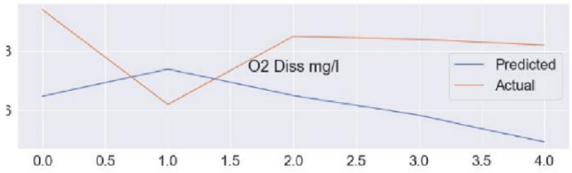

Figure 5: Time series of both actual and predicted Dissolved Oxygen in the next five years using MLP model.

Adam version of gradient descent is used as an optimizer. At each epoch, the model is evaluated on a validation set (test set). Mean Squared Error Losses for both training and testing, with respect to the LSTM model, are shown in figure 6.

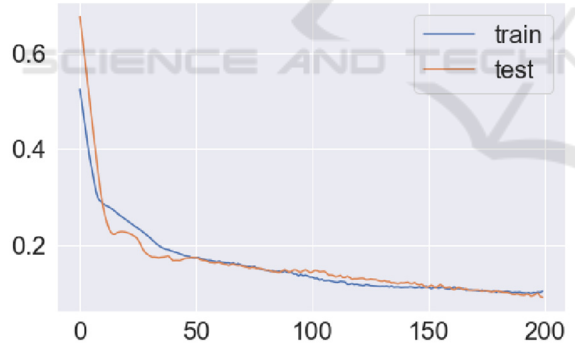

Figure 6: Train and Test losses for LSTM model.

Dissolved Oxygen is predicted for the next five years. Figure 7 displays the time series for both actual values (ground truth) and predictions for the next five years (the 1st year is 0 on the x-axis), using the LSTM model.

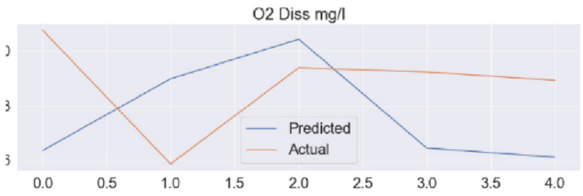

Figure 7: Time series of both actual and predicted Dissolved Oxygen in the next five years using LSTM model.

Figure 4 and figure 5 show that losses are decreasing considerably at an early time, proving that convergence is fast. It can be observed as well that LSTM loss converges faster than MLP loss.

As accurately. This is also observed from figure 5 and figure 7, which show that the accuracy of prediction decreases as the prediction time step increases. Moreover, the values of predicted and actual Dissolved Oxygen in LSTM at each year are closer to each other than those belonging to the MLP model, which yields errors (mse, mae and rmse) that are smaller.

4.2.2 Quantitative Results

To assess the prediction performance over the next five years, the two models are assessed concerning for three metrics, mean squared error (mse), mean absolute error (mae), and root mean squared error(rmse). Table 1 exhibits a summary of the results.

Table 1: Evaluation of the two models

Model	MSE	MAE	RMSE
MLP	6.2543	2.3911	2.5008
LSTM	**0.0924**	**0.2844**	**0.304**

The smaller the values of the aforementioned metrics, the better is the prediction accuracy of the model. MAE represents the average distance between the predicted value and the actual value. MSE is the average of the squared difference between predicted and actual values. RMSE is the square root of MSE, that gives errors in the same units as the variable itself.

From Table 1, it can be clearly shown that LSTM performs better than MLP in every single metric, as it reduces MSE, MAE, and RMSE respectively by 98.5%, 88% and 87%.

The LSTM model is capable of memorizing long-time steps, which allows for accurately predicting long-time durations.

5 CONCLUSIONS

The purpose of this study was to assess whether the two machine learning models are efficient in predicting the quality of Ibn Batouta dam water and to identify key water parameters allowing rapid and accurate monitoring of water quality (K. Chen et al. 2020) (M. Najafzadeh et al. 2020). Dam water quality prediction performance was comprehensively compared, and potential key water parameters were also defined and validated. Through this work, the main conclusions are:

1. LSTM performs better than MLP in each measurement because it is able to memorize longtime steps, which can accurately predict long durations.
2. The key parameters of water (Dissolved Oxygen) have been identified and validated by the two learning models MLP and LSTM.
3. Enriching the dataset with more experimental data can help in tuning the applied models and thus increasing the forecasting accuracy
4. Machine learning is recommended for future monitoring of dam water quality, as it could provide timely and accurate environmental alerts and further increase the efficiency of forecasting and decrease the cost of the dam forecast in the future monitoring of water quality.

ACKNOWLEDGEMENTS

The authors would like to thank all the collaborators within this work, from the Field sampling, laboratory analysis and writing manuscript team. El Khalil Cherif supported by FCT with the LARSyS - FCT Project UIDB/50009/2020 and by FCT project VOAMAIS (PTDC/EEI-AUT/31172/2017, 02/SAICT/2017/31172)

REFERENCES

Ali El Bilali et al. 2021. Groundwater quality forecasting using machine learning algorithms for irrigation purposes Agricultural Water Management Volume 245, February 28 2021, 106625 https://doi.org/10.1016/j.agwat.2020.106625

Amir Mosavi Water 2018Flood Prediction Using Machine Learning Models: Literature Review, 10, 1536; doi:10.3390/w10111536 www.mdpi.com/journal/water

Azzirgue E-M, and Salmoun F, 2019. Assessment of the Physico-Chemical Quality of Water of Oued Ouljat Echatt and Dam Ibn Batouta-Tangier. International Journal of Advances in Scientific Research and Engineering (ijasre). E-ISSN: 2454-8006. Volume 5, Issue 10 October – 2019.

Costabile, P.; Costanzo, C.; Macchione, F. A storm event watershed model for surface runoff based on 2D fully dynamic wave equations. Hydrol. Process. 2013, 27, 554–569.

Fernández-Pato, J.; Caviedes-Voullième, D.; GarcíaNavarro, P. Rainfall/runoff simulation with 2D full shallow-water equations: Sensitivity analysis and calibration of infiltration parameters. J. Hydrol. 2016, 536, 496–513.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Computation 9 (8): 1735–1780.

K. Chen et al. 2020 Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data / Water Research 171 (2020) 115454

Marius, P., Balas, V. E., Perescu-Popescu, L., Mastorakis, N. E. 2009. Multilayer perceptron and neural networks. WSEAS Transactions on Circuits and Systems 8(7).

Mohamed ACHMIT et al. 2017. Study of the physicochemical and bacteriological quality of waters of dam BAB LOUTA. International Journal of Innovation and Applied Studies ISSN 2028-9324 Vol. 20 No. Jul 4 . 2017, pp. 1246-1255.

M. Najafzadeh et al. 2020. Prediction of the five-day biochemical oxygen demand and chemical oxygen demand in natural streams using machine learning methods. Environ Monit Assess (2019) 191: 380 https://doi.org/10.1007/s10661-019-7446-8

Ouhmidou et al. 2015 Study of the physicochemical and bacteriological quality of the barrage Hassan Addakhil of Errachidia (Morocco) J. Mater. Environ. Sci. 6 (6) (2015) 1663-1671 ISSN: 2028-2508 CODEN: JMESCN

SETRAGEC 2018. EIE de l'AEP de la ville de Tanger et sa région à partir du barrage Ibn Battouta –conduite eau brute https://esa.afdb.org/sites/default/files/EIE%20MHARHAR-AEP%20Tanger.compressed.pdf

S. Jannicke Moe et al. 2019. Predicting Lake Quality for the Next Generation: Impacts of Catchment Management and Climatic Factors in a Probabilistic Model Framework. Water 2019, 11, 1767; doi:10.3390/w11091767, www.mdpi.com/journal/water

Umair Ahmed et al. 2019. Efficient Water Quality Prediction Using Supervised Machine Learning. Water 2019, 11, 2210; DOI: 10.3390/w11112210 www.mdpi.com/journal/water