An AI-based approach to the prediction of water points quality indicators for schistosomiasis prevention

Teegwende Zougmore
Université Alioune Diop de Bambey
Bambey, Senegal
teeg-wende@uadb.edu.sn

Bamba Gueye
Université Cheikh Anta Diop
Dakar, Senegal
bamba.gueye@ucad.edu.sn

Sadouanouan Malo
Université Nazi BONI
Nasso, Burkina Faso
sadouanouan@yahoo.fr

***Abstract*—We investigate the simultaneous daily prediction of the pH and temperature of a water point using AI-based methods. These parameters are part of the physicochemical parameters of surface water favoring the reproduction of parasitic worms responsible for Schistosomiasis. Wavelet Artificial Neural Network ($WANN$), Long Short Term Memory ($LSTM$) and Support Vector Regression (SVR) are the AI-based methods employed to build models with fifteen months collected data. They are evaluated through two metrics: root mean square ($RMSE$) and mean absolute error (MAE). The results show that in overall three methods give acceptable $RMSE$ which varies from 1.59 to 0.17. $WANN$ model shows the best performance with a $RMSE$ equals to 0.17 and a MAE equals to MAE 0.12 over $LSTM$ and SVR ones in forecasting parameters values one day ahead based on their two previous days observations.**

***Index Terms*—Machine Learning, Schistosomiasis, Water Quality Prediction, Wavelet transform**

I. Introduction

Schistosomiais is an infectious disease caused by a parasitic worm called schistosoma. Its transmission cycle starts by an infected person who releases parasite eggs through his urine or faeces in water points. The released eggs in some appropriate physical and chemical characteristics of water points hatch, penetrate and develop in snails until a stage of parasite. After snails release these parasites which enter in human body by skin contact where they develop until adult stage capable to lay eggs that human will release again causing the cycle to restart.

One can note through the transmission cycle that water points are the meeting place of all the actors. A relationship exists between Schistosomiasis infection and quality of water points. Water quality can be defined as the suitability of water for a particular application based on its chemical, biological and physical characteristics [1]. Predict water quality comes to forecast its variation trend at a certain time in the future.The main principle of Water Quality Prediction WQP is the estimation of one or more water parameters values in a short or long term time followed by an evaluation of set of conditions [2].

Many studies have addressed water quality prediction in the purpose to assess earlier pollution of water points which can cause water-related problems such as water-borne diseases and deaths of aquatic animals and so on [2] [3] [4] . Accuracy and long term forecasting of WQP have been addressed by many researchers. Some of them are reviewed in [3] and [4]. Although the satisfactory results reached with certain AI-based methods, there is still a need to investigate some methods in this study area. For example It has been stressed out in [3] that Support Vector Machine SVM performance has not yet explored in comparison to others AI techniques.

Few studies have addressed the modeling of water quality on some specific water-borne diseases especially Schistosomiasis. In [5], the authors considered three Machine Learning techniques which are SVM, Random Forest (RF) and Artificial Neural Network (ANN) to build a earlier detection tool. They aim with this tool to assess the suitability of water points for schistosoma eggs maturation and intermediate hosts (snails) development. Based on fifteen days collected data, the tool has been trained on different sliding windows namely 1 hour, 2 hours, 3 hours and 6 hours. SVM performs well over the two other algorithms in all sliding windows. We note that the forecasting horizon doesn't reach one day.

We investigate AI methods (machine learning and deep learning) in this study which can permit to build models capable to forecasting one day ahead PH and temperature of a water point. We explore one day ahead horizon because the result of the water quality prediction is going to combine with a mathematical model which can provide daily evolution of snails and parasites densities. We have described the fusion conceptual framework in this work [6]. The objectives of our study here are twofold. First is to investigate the forecasting of water quality favorable to Schistosomiasis transmission at one day ahead. Second we address one of the recommendations enumerated in [3] which is to pay more attention in comparison of single AI methods to hybrid ones such as $WANN$.

The rest of the paper is organized as follows. Section II gives the principle of different AI methods employed. In section III, we develop the different steps of model building. Section IV presents the study area, the collected data, the evaluation metrics and results. We draw a conclusion and give perspectives in section V.

II. Background on used AI methods

A. Support Vector Regression

SVM is a machine learning which finds the maximum margin separating hyperplane which correctly classifies as possible many training points of a dataset. SVM can be employed for classification and regression problems. Its variant dedicated to regression problems is called support vector regression (SVR). The modification concerns the constraints. $y^{(i)}$ and his predicted value $\langle w, x^{(i)} \rangle$ must be less to a certain value ϵ The optimization problem addressed by SVR is expressed as follow :

$$\begin{aligned} &\text{min } w \in \Re^p, b \in \Re, \epsilon \in \Re^n \ \tfrac{1}{2} \ \|w\|_2^2 \ + \ C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \\ &\text{such that } y^{(i)} \ - \ \langle w, x^{(i)} \rangle - b \ \leq \ \epsilon + \xi_i \\ &\ and \ \langle w, x^{(i)} \rangle + b \ - y^{(i)} \leq \epsilon + \xi_i^* \quad \forall i \in \ \{1, ...n\} \end{aligned} \tag{1}$$

Where w represents the normal vector to hyperplane and $y^{(i)}$ is the label (-1 or 1) of an i^{th} observation. b is a scalar which determines the axis intercepts. ξ_i are slack variables which makes the margin to be soft (i-e allow some errors during training phase). C is a regularization parameter. n and p represent respectively the number of samples and predictors of the dataset.

B. Long Short Term Memory

Long Short-Term Memory (LSTM) is a specific recurrent neural network (RNN) architecture that was designed to model temporal sequences and their long-range dependencies more accurately than conventional RNNs [7]. The specificity of $LTSM$ resides on the structure of neurons. Neurons are designated as blocks. Each block is constituted a cell and three gates which control the cell. The presence of the cell permits to handle long term dependency and vanishing/exploding problems encountered with RNN architecture. For more details, one can refer to [8].

C. Wavelet Artificial Neural Network

$WANN$ is an hybrid technique which employs wavelet transform and artificial neural network to make time series forecasting.

Based on definition given by [9], we can stress out that WT is a method which can be used to convert a temporal signal into another form which makes certain features (trend, noise, changes, etc) more amenable to study. There are two major transform continuous wavelet transform (CWT) and discrete wavelet transform (DWT).

DWT fits with time series as they use discrete time. It decomposes time series into sub-time series. One of them represent the slow changing of the time series and the others designate the fast changing. The slow changing or the trend are characterised by approximations coefficients and the fast changing (variations) by details coefficients. These sub-times serve as inputs to an ANN architecture.

ANN is a computational system of interconnected nodes inspired by biological neural networks of human brain. A typical ANN contains a large number of nodes arranged in a series of different layers : input layer, hidden layer and output layer as depicted in Fig. 1.

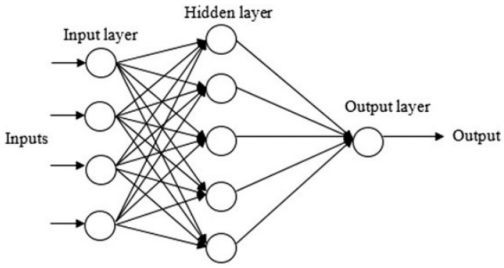


Fig. 1: ANN architecture with one hidden layer [10].

Fig.2 illustrates the combination of DWT and ANN. The ANN architecture employed here is feed-forward neural network with back-propagation.

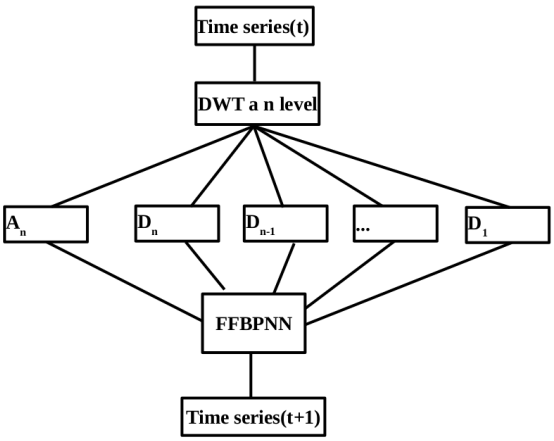


Fig. 2: WANN architecture

A_n represents the approximation coefficients and D_i with i ranging from 1 to n, represent details coefficients. The level of decomposition is computed following this equation.

$$level = int(\log_N) \tag{2}$$

with N represents the number of samples in the dataset. A trial and error method is applied to determine the fittest wavelet between. We have considered daubechies8 and haar in this study. They are adapted to time series variance analysis.

III. Models building

Multiple parallel input and multi-step output is the scheme of forecasting investigated. It consists in the case of multivariate time series to estimate in a certain future the values of each series separately. The different steps in building such model are described in the following sections.

A. Data pre-processing

This step consists in eliminating negative values and values which are out of the ranges [0-14] for PH and [-55 +125] for temperature. The ranges are defined based on the sensors capacity which is indicated in the specifications of the sensors.

In addition to eliminate values based on conditions mentioned above, data are re-sampled in one day frequency. Table I and Table II show respectively an overview of initial data and an overview of daily average data.

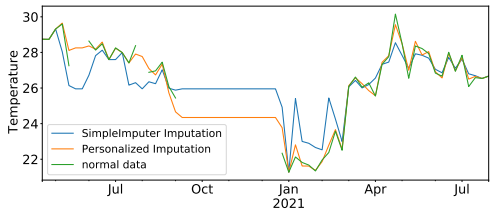
TABLE I: An overview of raw dataset

DATE	TEMP	PH
2020-04-14 17:48:00	30.00	5.18
2020-04-14 17:53:32	29.94	5.20
2020-04-14 17:59:04	29.88	5.20
2020-04-14 18:04:36	29.88	5.18
2020-04-14 18:10:08	29.88	5.20

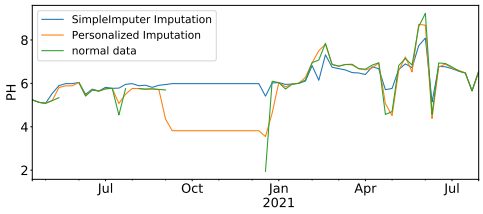
TABLE II: An overview of daily averaged data

DATE	TEMP	PH
2020-04-15	28.863594	5.284009
2020-04-16	27.575000	5.334954
2020-04-17	27.875117	5.232394
2020-04-18	29.023423	5.221081
2020-04-19	29.441629	5.204389

Two methods are performed to handle missing values. We tried mean imputation through SimpleImputer class from scikit-learn [11]. The second method is designed by us. It consists to fill missing values by an average value calculated with the two values separating each missing period in the series. Fig.3 shows three curves. one in green for represents data with missing values. The two others (blue for mean imputation and orange for personalized imputation) are the estimations of missing values done by the imputations methods.



(a) Temperature Data



(b) PH Data

Fig. 3: Results imputation methods

To better appreciate the mentioned imputation methods, let us consider Fig.3a. Curves must be confused in the period January to April 2021 for there is no missing values. But one can notice that normal curve and personalized imputation curve tend to be confused over mean imputation curve. The estimations done by personalized imputation method is more acceptable than mean imputation. We have preferred this imputation method in the study.

B. Data preparation

In this step, data are transformed from time series into a supervised learning format. It means that the lagged observations of different parameters are considered as independent variables or predictors. And the ahead observations are considered like dependent variables. Table.III shows a few rows of how the data are transformed into supervised learning format.

TABLE III: series transformed into supervised learning format

PH $(t-2)$	TEMP $(t-2)$	PH $(t-1)$	TEMP $(t-1)$	PH (t)	TEMP (t)	PH $(t+1)$	TEMP $(t+1)$
5.180	29.663	5.284	28.864	5.335	27.575	5.232	27.875
5.284	28.864	5.335	27.575	5.232	27.875	5.221	29.023
5.335	27.575	5.232	27.875	5.221	29.023	5.204	29.442

param(t-x) with param being {PH,TEMP} and x ranges from 2 to 1 are the lagged observations (of two and one days ago) to use to predict two days ahead forecast observations of PH and temperature represented by param(t) and param(t+x).

C. Data splitting

Data are subdivided in three parts which are :

- Training set : from April 2020 to March 2021
- Validation set : from March 2021 to April 2021
- Test set : from May 2021 to July 2021.

D. Models definition

Three layers have been adopted for $LSTM$ and $WANN$.One input, one hidden and one output layer. The number of nodes for input and output layers are determined dynamically through forecast horizon which ranges from 1 to 7 and lag length which ranges also from 1 to 7. The number of series of the dataset is also considered. It equals to 2 as we deal with two parameters. Especially for $WANN$, the decomposition level is considered. It is determined by equation 2. We can summary the determination of the number of nodes for input and output layers in the table IV.

TABLE IV: Number of nodes of input and output layers

Method	Input Layer	Output Layer
$WANN$	$(i+1)*n$	$H*n$
$LSTM$	$L*n$	$H*n$

H represents the forecasting horizon. L represents the lag length. n represents the number of series and i is the level of decomposition. The number of epochs and nodes of hidden layers are determined by trial and error. Both equal 100 in this study.

SVR is wrapped in a specific class of scikit-learn MultiOutputRegressor which gives the ability to SVR to

predict separately the outputs.

IV. Evaluation

A. Study area and data

1) Study area: An assembled device constituted of an Arduino microcontroller and low cost sensors bought on online market especially DFRobot and Aliexpress has been placed in a backwater to measure its characteristics namely pH, electrical conductivity, turbidity, dissolved oxygen, temperature, flow and total dissolved solids. The backwater located in Panamasso which is a village of the district of Houet in Burkina Faso. Its coordinates are latitude 11°23'0"North and longitude 4°12'0"West.

2) Data: Data have been collected regularly in five minutes frequency from April 2020 to July 2021. But we stress out that the device has encountered sometimes some dysfunctions which cause a data missing of during certain periods. The collected data considered concern values of temperature and pH parameters. Collected data are stored into an IoT platform (thingspeak.com) accessible at this url:https://thingspeak.com/channels/963425.

B. Background metrics

Root Mean Squared Error ($RMSE$) and Mean Absolute Error (MAE) are the metrics used to assess the models. Both indicate how far the values forecasted by the model are from expected ones. They are expressed in the same unit as the target value that is being forecasted. Also they are relative to the dataset. Lower they are, better the model is. They are calculated as follows :

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \widehat{y_i})^2} \tag{3}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \widehat{y_i}| \tag{4}$$

with m as the size of dataset. y_i corresponds to the target value and $\widehat{y_i}$ is the forecasted one. $\overline{y}$ is the target mean value.

$RMSE$ penalizes a model's errors, therefore more attention is paid to this metric. If two models have the same $RMSE$, it is MAE which is considered to decide between them.

C. Results and discussion

The number of models built for each method is 49. They have been evaluated ten times. We obtain 49 models for each method due to the fact that we explore for a range of 7 lagged observations, a range of 7 forecasting horizons. To distinguish the models, we used a notation taking account the lag and horizon numbers. The notation is Lx_Hy with x and y ranges $[1-7]$. If $x = 3$ and $y = 2$ for example, it means that the model has used two lagged days observations to forecast two days ahead values.

For the appreciation of the methods, we proceed as follows:

- 1 for each method, we group by the models obtained.
- 2 for each model we calculated the mean value of each metric obtained during the ten times evaluation
- 3 we sort the result of stage 2 according to the minimum $RMSE$

The Table V presents the best model of each method.

TABLE V: Evaluation of metrics

Method	RMSE	MAE	Model
$WANN$	0.17	0.12	$L2\_H1$
$LSTM$	1.15	0.72	$L2\_H1$
SVR	1.24	0.88	$L3\_H1$

Considering the values of metrics in Table V, the $WANN$ model $L2\_H1$ performs well over the others models. To confirm this result we plot the average $RMSE$ of these best models present in Table V. We obtain the graphics showed in Fig.4. The graphics indicate that the $WANN$ model $L2\_H1$ performs well over the $LSTM$ model $L2\_H1$ and SVR model $L3\_H1$ during ten evaluations. We can notice that curve representing $WANN$ best model is positioned lower than the two others.

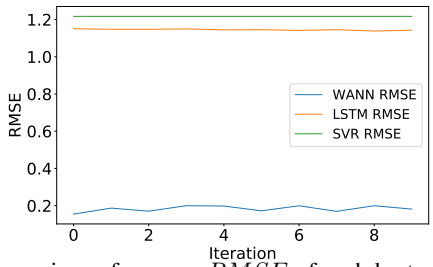
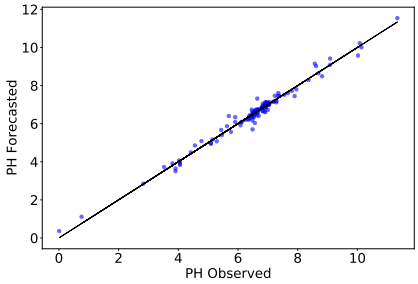


Fig. 4: Comparison of average $RMSE$ of each best model obtained with each method.

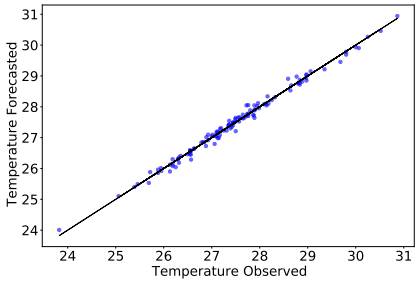
In Fig.5, we present a visual representation of observed PH and Temperature values against forecasted ones by the $WANN$ model $L2\_H1$. Same visual representations are presented in Fig.6 and Fig.7 where the forecasted values are obtained respectively by $LSTM$ model $L2\_H1$ and SVR model $L3\_H1$. With these plots, we can check how much the forecasted values are far from observed ones. An ideal situation is to get all the points in the straight line. The model which is closed to this is $WANN$ model $L2\_H1$. Observed values of PH and Temperature are almost confused to forecasted ones in Fig.5a and in Fig.5b. This not the case in Fig.6a, Fig.6b, Fig.7a Fig.7b.

Wavelet analysis combined to ANN architecture performs well based on metrics $RMSE = 0.17$ and $MAE = 0.12$ over SVM and $LSTM$. The reachable performance with this method indicates that is possible to forecast one day with an error less than 0.2 for each value of PH and temperature.

SVR doesn't support inherently multi-step output. This may explain its relative bad performance over $WANN$ and $LSTM$. $LSTM$ handles well temporal sequence. Its performance is surprising but one must note that it is a deep learning technique which performs well when the volume of

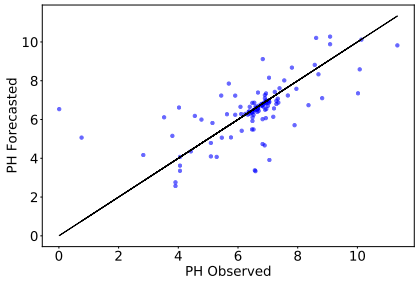


(a) Predictions vs Observations PH

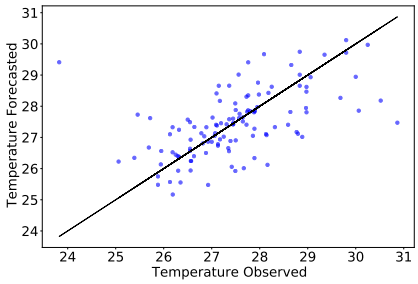


(b) Predictions Vs observations - Temperature

Fig. 5: WANN best model ($L2\_H1$)



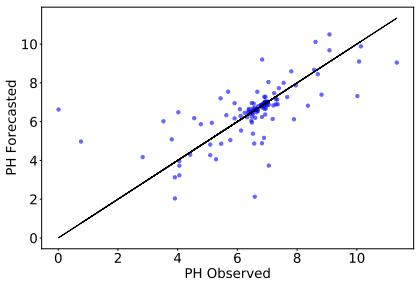
(a) Predictions vs Observations - PH



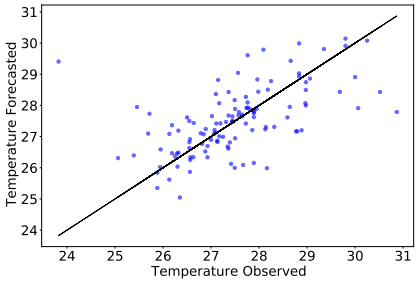
(b) Predictions Vs observations - Temperature

Fig. 6: LSTM best model ($L2\_H1$)

data is huge. The quantity of data in this study could explain its relative bad performance compared to $WANN$. $WANN$ is an hybrid which is presented as an efficient technique in the literature. The result achieved here is not contradictory. It capacity to transform a time series into its trend and variations parts and consider these parts as ANN inputs could explain



(a) Predictions vs Observations - PH



(b) Predictions Vs observations - Temperature

Fig. 7: SVR best model ($L3\_H1$)

the good performance achieved with this technique.

V. Conclusion

AI-based methods can be employed to build efficient water quality prediction models. We have investigated in this study the performance of three methods namely SVR, $LSTM$ and $WANN$. It is on data collected from an endemic place of Schistosomiasis that the methods have been applied to build models. We have considered basic configuration of each method without tuning their hyperparameters. $WANN$ which is a hybrid model outperforms SVR and $LTSM$ which are single AI methods.

$WANN$ gives good $RMSE$ in horizon of one day simultaneous forecasting of PH and Temperature. This is result is satisfactory for It then possible with $WANN$ to assess accurately one day in advance the appropriate conditions of the reproduction of parasitic worms responsible for Schistosomiasis.

Water quality for Schistosomiasis entities development concerns also others parameters such as electrical conductivity, dissolved oxygen and total dissolved solids. An extension of this study taking in account all these parameters is a future work to do. Also investigate the three methods employed here with more data can be interesting for we could observe their behavior on large dataset of water quality.

References

[1] M. Pule, A. Yahya, J. Chuma, M. Pule, A. Yahya, and J. Chuma, "Wireless sensor networks: A survey on monitoring water quality," *Journal of applied research and technology*, vol. 15, no. 6, pp. 562–570, 2017.

[2] T. Jin, S. Cai, D. Jiang, and J. Liu, "A data-driven model for real-time water quality prediction and early warning by an integration method," *Environmental Science and Pollution Research*, vol. 26, no. 29, pp. 30 374–30 385.

[3] T. Rajaee, S. Khani, and M. Ravansalar, "Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: A review," *Chemometrics and Intelligent Laboratory Systems*, vol. 200, p. 103978.

[4] Y. Chen, L. Song, Y. Liu, L. Yang, and D. Li, "A review of the artificial neural network models for water quality prediction," *Applied Sciences*, vol. 10, no. 17, p. 5776.

[5] B. Kassé, B. Gueye, M. Diallo, F. Santatra, and H. Elbiaze, "IoT based Schistosomiasis Monitoring for More Efficient Disease Prediction and Control Model," in *2019 IEEE Sensors Applications Symposium (SAS)*, Mar. 2019, pp. 1–6.

[6] T. Zougmore, S. Malo, B. Gueye, and S. Ouaro, "Toward a data fusion based framework to predict schistosomiasis infection," in *2020 IEEE 2nd International Conference on Smart Cities and Communities (SCCIC)*, 2020, pp. 1–8.

[7] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," 2014.

[8] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[9] P. S. Addison, *The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance*, second edition ed. CRC Press, 2017.

[10] M. Ravansalar, T. Rajaee, and M. Ergil, "Prediction of dissolved oxygen in river calder by noise elimination time series using wavelet transform," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 28, no. 4, pp. 689–706.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.