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

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

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

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

Abstract

We investigate the simultaneous daily forecasting of pH, temperature, dissolved oxygen, and electrical conductivity using AI-based methods. These physicochemical parameters can be retrieved from surface water and favor the reproduction of parasitic worms responsible for Schistosomiasis. Wavelet Artificial Neural Network ($WANN$), Long Short Term Memory ($LSTM$), and Support Vector Regression (SVR) are used AI-based methods to build models with fifteen months of collected raw datasets. They are evaluated through two metrics, such as root-mean-square ($RMSE$) and mean absolute error (MAE). The built models take as inputs the physicochemical parameters values observed the last two days and provide as outputs the physicochemical parameters values expected the next day. Overall, the results show that the three methods perform well. The most efficient according to the metrics is the WANN-based model which shows a RMSE of 0.07, 0.13, 0.09, and 9.79 in forecasting respectively pH, temperature, dissolved oxygen, and electrical conductivity.

1 Introduction

Schistosomiasis is an infectious disease caused by a parasitic worm called Schistosoma. The World Health Organization (WHO) estimated in 2021 that at least 251.4 million people needed preventive treatment against it. At least 90% of these people lived in Africa 1. Tropical and intertropical regions are the preferred areas of prevalence of Schistosomiasis. It constitutes the second parasitic endemic in the world after malaria 2.

Its transmission cycle starts with 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 the parasite. After snails release these parasites, which enter the human body by skin contact where they develop until the adult stage capable to lay eggs that humans will release again, causing the cycle to restart.

The quality of water points influences the biological cycle of snails and parasites [3]. Water quality can be defined as the suitability of water for a particular application based on its chemical, biological, and physical characteristics [4]. Predicting 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

5th Deep Learning Indaba Conference (DLI 2023).

one or more water parameter values in a short or long-term time, followed by an evaluation of a set of conditions [5].

Many studies have addressed water quality prediction with 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 [5; 6]. Accuracy and long-term forecasting of WQP have been addressed by many researchers. Some of them are reviewed in [6] and [7]. Although satisfactory results were reached with certain AI-based methods, there is still a need to investigate some methods in this study area. For instance, it has been stressed in [6] that Support Vector Machine SVM performance has not yet been explored in comparison to other AI techniques.

Few studies have addressed water modeling quality on some specific water-borne diseases, especially Schistosomiasis. In [8], the authors considered three Machine Learning techniques which are SVM, Random Forest (RF), and Artificial Neural Network (ANN) to build an earlier detection tool. They aim to assess the suitability of water points for Schistosoma egg maturation and intermediate hosts (snails) development. Based on fifteen days of 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 the use of AI methods to build models that can forecast water quality parameters, such as pH, temperature, dissolved oxygen, and electrical conductivity, of a water point one day in advance. We explore one day ahead horizon since the water quality prediction result is going to combine with a mathematical model which can provide daily evolution of snails and parasite's densities. We have described the fusion conceptual framework in a previous work [9].

The objectives of our study here are twofold. Firstly, we investigate the forecasting of water quality favorable to Schistosomiasis transmission one day ahead. Secondly, we address one of the recommendations enumerated in [6] which is to pay more attention to the comparison of single AI methods to hybrid ones such as $WANN$.

The rest of the paper is organized as follows. Section 2 gives an overview of the proposed system's structure and describes the principle of different AI methods explored. In section 3, we present the determination of hyperparameters and different steps followed in model building. Section 4 presents our experimental setup and results. Finally, we conclude and give perspectives in section 5.

2 Structure of proposed system and used AI methods backgrounds

Figure 1 indicates the general structure of the proposed system, constituted of an IoT-based system for capturing data and an AI-based model which relies on previous data of some physicochemical parameters to forecast their future values.

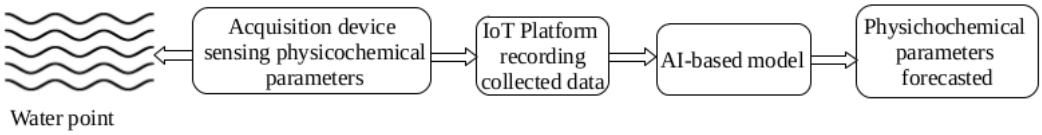

Figure 1: Structure of proposed system

We focus on this paper on the AI-based model building. We have explored some AI methods in that purpose, namely support vector regression (SVR), long-term short memory ($LSTM$), and wavelet artificial neural networks ($WANN$). SVR is a variant of support vector machine (SVM) dedicated to regression problems [10]. It aims to find an approximation function that allows for the estimation of target values while keeping the deviation within a specified tolerance level ϵ.

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 [11]. The specificity of $LTSM$ resides in the structure of neurons. Neurons are designated as blocks. Each block is constituted of a cell and three gates that 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 [12].

$WANN$ is a hybrid technique which employs wavelet transform (WT) and artificial neural network (ANN). Based on the definition given in [13], we can stress that WT is a method that 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 transforms, 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 represents the slow changing of the time series and the others designate the fast changing. The slow changing or the trend is characterized 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 the biological neural networks of the human brain. A typical ANN contains numerous nodes arranged in a series of different layers: input layer, hidden layer, and output layer.

3 Models training steps and hyperparameters determination

Multiple parallel input and multistep output is the scheme of forecasting investigated. It consists in the case of multivariate time series to estimate simultaneously in certain future the values of each series separately.

Various steps have been followed in model training. Firstly, data have been preprocessed by removing negative and out-of-range values. Secondly, data have been resampled in one-day frequency. To handle missing values, we have designated thirdly an algorithm that consists of filling missing values with an average value calculated with the values separating each missing period in the series. Fourthly, data are transformed from time series into cross-sectional data. It means that the lagged observations of different parameters are considered independent variables or predictors. And the ahead observations are considered dependent variables. Finally, data have been split into training set (from April 2020 to March 2021); validation set (from March 2021 to April 2021); test set (from May 2021 to July 2021).

The training models require some hyperparameters. After trial and error, 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 is 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 4 as we deal with four parameters. Especially for $WANN$, the decomposition level has to be considered according to the number of dataset samples (denoted N). It is determined by equation 1.

$$level = int(\log_N) \tag{1}$$

The number of nodes for input and output layers of $WANN$ and $LSTM$ is summarized in table 1.

Table 1: Number of nodes of input and output layers

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

Based on table 1, H, L,n, i represents respectively the forecasting horizon, the lag length, the number of series, 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 algorithm's hyperparameters concern the used kernel, the width of this kernel (γ), the ϵ tube's width, and the regularization parameter C. These hyperparameters are presented in Table 2.

Table 2: SVR hyperparameters

Kernel	ϵ	C	Γ
Gaussian	0.001	5	0.001

The value of C was determined after testing different values such as 0.01,0.1,1,5,10, and 100. SVR does not support multi-output directly. It has then been wrapped in a specific class of scikit-learn MultiOutputRegressor which gives it the ability to predict separately many outputs [14].

4 Evaluation

4.1 Study area and data

An assembled device constituted of an $Arduino$ microcontroller and low-cost sensors has been placed in a backwater to measure its characteristics namely pH, Temp, EC, DO, turbidity, flow, and total dissolved solids. The backwater is 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.

The raw datasets have been collected regularly in five minutes frequency from April 2020 to July 2021. But we stress that the device has encountered sometimes some dysfunctions which cause data missing during certain periods. In this study, we consider only values of pH, Temp, DO, and EC for as we have mentioned due to some dysfunctions, some sensors haven't provided good data. Collected data are stored in an IoT platform (thingspeak.com) accessible at this link: https://thingspeak.com/channels/963425.

4.2 Results and discussion

The number of models built for each method is 49. We obtain 49 models for each method because we explore lag lengths and forecast horizons ranging from 1 to 7 days. To distinguish the models, we used a notation taking into account the lag length and forecast horizon. The notation is $METHODLx\_Hy$ with x and y ranges $[1-7]$. $METHOD$ being $WANN$ or $LSTM$ or SVR. If $x = 3$ and $y = 2$, for example, it means that the model has used three lagged days observations to forecast two days values.

Root Mean Squared Error ($RMSE$) and Mean Absolute Error (MAE) are used metrics to assess the models. $RMSE$ penalizes a model's errors, therefore more attention is paid to this metric. If two models have the same $RMSE$, MAE is considered to decide between them. The reader can refer to [15] for more details.

For methods assessment, we select for each method the model with the minimum $RMSE$ among the models obtained. We note that each method, it is the model L_2_H1 which presents the minimum $RMSE$. Afterward, we compare these three models as presented in Table 3 to retain the one which performs well over the others.

Table 3: Models performances

Metric	PH	Temp	OD	EC	Model
RMSE	0.07	0.13	0.09	9.79	WANN L2_H1
MAE	0.05	0.06	0.06	7.15	
RMSE	1.04	1.00	0.34	37.40	LSTM L2_H1
MAE	0.62	0.73	0.22	27.63	
RMSE	0.91	1.06	0.02	11.40	SVR L2_H1
MAE	0.53	0.71	0.01	8.45	

Considering the values of metrics in Table 3, the $WANN$ model $L2\_H1$ performs well over the other models.

SVR does not support inherently multistep 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 that performs well when the volume of data is huge. The quantity of data in this study could explain its relative bad performance compared to $WANN$. $WANN$ is a hybrid method that is presented as an efficient technique in the literature [6]. The result achieved here is not contradictory. Its capacity to transform a time series into its trend and variations parts and consider these parts as ANN inputs could explain the good performance achieved with this technique.

The performance of the models developed in this study is consistent with the findings of previous research, which has shown that AI-based methods are effective for predicting water quality [6]. The models developed in this study can simultaneously predict pH, temperature, dissolved oxygen, and electrical conductivity. This differs from previous models, which have typically focused on predicting individual physicochemical parameters [6; 7].

5 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. The scheme of forecasting is multiple parallel inputs and multistep output. We have determined the hyperparameters of the investigated methods by trial and error. $WANN$ which is a hybrid model outperforms SVR and $LTSM$ which are single AI methods.

$WANN$ based model gives good $RMSE$ in the horizon of one-day simultaneous forecasting of pH, Temperature, dissolved oxygen, and electrical conductivity which are respectively 0.07, 0.13, 0.09, and 9.79. The obtained result is satisfactory, since it is possible with a $WANN$-based model to assess accurately one day ahead the appropriate conditions that influence the biological cycle of snails and parasitic worms responsible for Schistosomiasis.

Nevertheless, we have not considered some physicochemical parameters which can influence also the biological snail's cycle and parasites such as water flow and turbidity due to some dysfunctions of a couple of device acquisitions. It would be interesting to consider these parameters and investigate the behaviors of the models when the number of physicochemical parameters increases.

References

[1] O. mondiale de la Santé, W. H. Organization, *et al.*, "Schistosomiasis and soiltransmitted helminthiases: progress report, 2021–schistosomiase et géohelminthiases: rapport de situation, 2021," *Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire*, vol. 97, no. 48, pp. 621–631, 2022.

[2] P. Aubry and B. Gaüzère, "Schistosomoses ou bilharzioses," *Médecine Tropicale*, vol. 8.

[3] S. Bakhoum, R. A. Ndione, C. J. E. Haggerty, C. Wolfe, S. Sow, C. T. Ba, G. Riveau, and R. R. Jason, "Influence des paramètres physico-chimiques sur la répartition spatiale des mollusques hôtes intermédiaires des schistosomes humains dans le delta du fleuve sénégal," *Médecine et Santé Tropicales*, vol. 29, no. 1, pp. 61–67.

[4] 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.

[5] 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. 30374–30385.

[6] 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.

[7] 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.

[8] 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)*, pp. 1–6, Mar. 2019.

[9] 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)*, pp. 1–8, 2020.

[10] M. Awad and R. Khanna, "Support Vector Regression," in *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers* (M. Awad and R. Khanna, eds.), pp. 67–80, Berkeley, CA: Apress, 2015.

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

[12] 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.

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

[14] 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.

[15] M. Naser and A. H. Alavi, "Error metrics and performance fitness indicators for artificial intelligence and machine learning in engineering and sciences," *Architecture, Structures and Construction*, pp. 1–19, 2021.