

# *Predicting And Analyzing Urban Water Quality Using IBM Watson Machine Learning Service*

## 1. INTRODUCTION

### OVERVIEW

**Water** (chemical formula **H<sub>2</sub>O**) is an [inorganic](#), transparent, tasteless, odorless, and [nearly colorless](#) [HYPERLINK "https://en.wikipedia.org/wiki/Color\\_of\\_water"](https://en.wikipedia.org/wiki/Color_of_water) [HYPERLINK "https://en.wikipedia.org/wiki/Color\\_of\\_water"](https://en.wikipedia.org/wiki/Color_of_water) [rless chemical substance](#), which is the main constituent of [Earth's hydrosphere](#) and the [fluids](#) of all known living organisms (in which it acts as a [solvent](#)). It is vital for all known forms of [life](#), even though it provides neither [food, energy](#), nor organic [micronutrients](#). Its [chemical formula](#), H<sub>2</sub>O, indicates that each of its [molecules](#) contains one [oxygen](#) and two [hydrogen atoms](#), connected by [covalent bonds](#). The hydrogen atoms are attached to the oxygen atom at an angle of 104.45°. "Water" is also the name of the liquid state of H<sub>2</sub>O at [standard temperature and pressure](#).

A number of natural states of water exist. It forms [precipitation](#) in the form of rain and [aerosols](#) in the form of fog. [Clouds](#) consist of suspended droplets of water and [ice](#), its solid state. When finely divided, [crystalline](#) ice may precipitate in the form of [snow](#). The gaseous state of water is [steam](#) or [water vapor](#).

Water covers about 71% of the Earth's surface, mostly in seas and oceans (about 96.5%). Small portions of water occur as [groundwater](#) (1.7%), in the [glaciers](#) and the [ice caps](#) of [Antarctica](#) and [Greenland](#) (1.7%), and in the air as [vapor](#), clouds (consisting of ice and liquid water suspended in air), and [precipitation](#) (0.001%). Water moves continually through the [water cycle](#) of [evaporation](#), [transpiration](#) ([evapotranspiration](#)), [condensation](#), precipitation, and [runoff](#), usually reaching the sea.

The aim of this study is the prediction of water quality components using artificial intelligence (AI) techniques including MLP, SVM, and group method of data handling (GMDH). Therefore, in the first part of this section, the studied area is introduced and then ranges of measured water quality components are presented. Overviews on applied AI models are then presented.

## PROPOSED SYSTEM

Water is considered as a vital resource that affects various aspects of human health and lives. The quality of water is a major concern for people living in urban areas. Quality of water serves as a powerful environmental determinant and a foundation for the prevention and control of waterborne diseases. However predicting the urban water quality is a challenging task since the water quality varies in urban spaces non-linearly and depends on multiple factors, such as meteorology, water usage patterns, and land uses, so this project aims at building a Machine Learning (ML) model to Predict Water Quality by considering all water quality standard indicators.

The main purpose of the project is to analyse and predict the urban water quality in which it consists of different florides and chlorides. These florides and chlorides should be in smaller portion. In order to find these quantities we use methods like

### *Temperature testing*

Testing the temperature helps determine the rate of biochemical reaction in an aquatic environment and indeed whether they are able to occur at all. If the water temperature is too elevated, this can limit the water's ability to hold oxygen and decrease organisms' capacity to resist particular pollutants.

### *pH testing*

Measures the acidity of water. Most aquatic organisms are only able to survive within a pH range of 6 to 8.

### *Chloride test*

Chloride is usually present in fresh and salt water. However, its levels can be exacerbated as a result of minerals dissolving and industrial pollution.

### *Turbidity test*

Measures the amount of particulate matter that is suspended in the water, or more simply, how clear the water is. If high levels of turbidity are present, photosynthesis is affected as light is unable to penetrate, increasing water temperature.

### *Dissolved oxygen test*

Measures the amount of oxygen dissolved in water. Without this, aquatic life is unable to conduct cellular respiration and is thus a key indicator of water health.

### *Nitrate and Phosphate test*

The presence of these essential nutrients is a good indicator of strong plant life. However, the addition of artificial nitrates and phosphates through detergents, fertilisers or sewage can be harmful and result in eutrophication, generally in the form of unwanted algal blooms.

### *Electrical conductivity test*

Estimates the total amount of solids dissolved in the water. This can be a good indicator of the level of salinity.

### *Salinity testing*

Measures the total of all non-carbonate salts dissolved in water. Measuring groundwater salinity indicates how salty your topsoil may become if the watertable rises.

So to know the quality of water we use the above traditional methods. But in this we will predict the quality of water by machining learning model.

## **2. LITERATURE SURVEY**

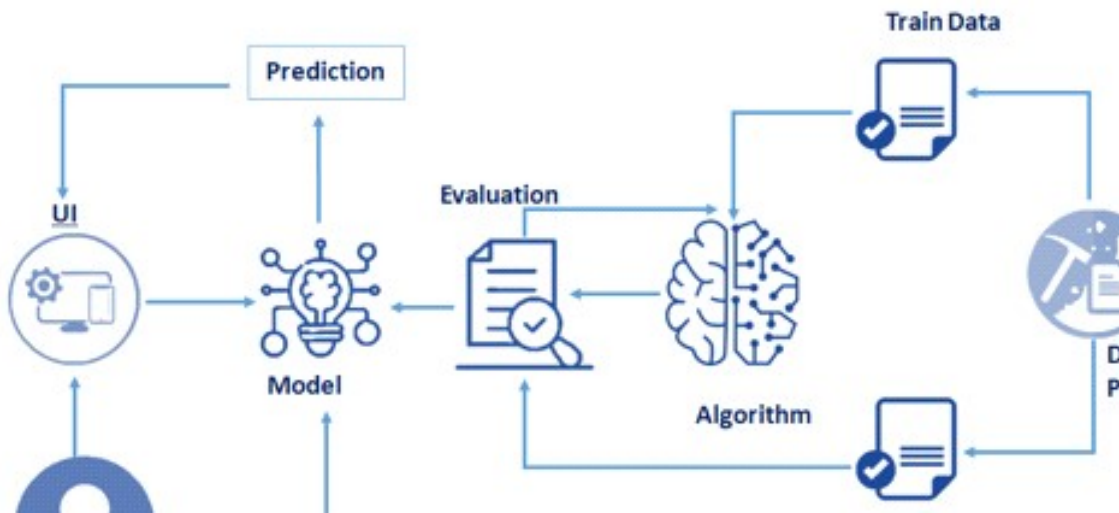
Groundwater quality parameters in selected points of rural, sub urban and urban areas. But none have exclusively studied the confluence zone. In this backdrop the investigator of this project has proposed to undertake this study with UGC support which is locally relevant and has lab to land orientation as well. Since the levels of certain physico - chemical parameters like pH, alkalinity, BOD, COD, total hardness (Ca,Mg,Chlorides, $\text{HCO}_3^-$ ,  $\text{CO}_3^{2-}$ ),  $\text{NO}_3^-$ ,  $\text{PO}_4^{3-}$ , Na, K and total dissolved solids found in ground water determine its behaviour as well as quality, it has been proposed to analyse ground water in the five slum areas separately so as to find reasons for the ground water problems in the said area.

### *Data collection*

ML depends heavily on data, without data, it is impossible for an "AI" to learn. It is the most crucial aspect that makes algorithm training possible. In Machine Learning projects, we need a training data set. It is the actual data set used to train the model for performing various actions.

### 3. THEORETICAL ANALYSIS

#### 3.1 BLOCK DIAGRAM



#### *Water Quality Index Calculation*

To measure water quality, WQI is used to be calculated using various parameters that significantly affect WQ [40–42]. In this study, a published dataset is considered to test the proposed model, and seven significant water quality parameters are included. The WQI has been calculated using the following formula:

Where  $N$  is the total number of parameters included in the WQI calculations  $q_i$  is the quality rating scale for each parameter  $i$  calculated by equation (2) below, and  $w_i$  is the unit weight for each parameter calculated by equation

Where  $x_i$  is the measured value of parameter  $i$  in the tested water samples  $x_i$  is the ideal value of parameter  $i$  in pure water (0 for all parameters except DO = 14.6 mg/l and pH = 7.0), and  $S_i$  is the recommended standard value of parameter  $i$

#### *Z-Score Normalization Method*

Normalization is a way to simplify calculations. It is a dimensional expression transformed into a nondimensional expression and becomes a scalar. Z-score normalization (or normalization score) is a normalization method used to

normalize parameters by using the mean ( $\mu$ ) and standard deviation ( $\sigma$ ) values of the tested data. It can be calculated as follows:

$$Z\text{-score} =$$

Permissible limits of the parameters used in calculating WQI

Parameters	Permissible limits
Dissolved oxygen, mg/l	10
pH	8.5
Conductivity, $\mu\text{S/cm}$	1000
Biological oxygen demand, mg/l	5
Nitrate, mg/l	45
Fecal coliform, Cfu/100 ml	100
Total coliform, Cfu/100 ml	1000

### 3.2 Prerequisites

Hardware Requirements:

- 2GB RAM or Above
- Dual Core processor or Above
- Internet Connection

Software Requirements:

- Anaconda Navigator
- Python Packages
- IBM Watson Studio

#### **4. EXPERIMENTAL INVESTIGATION**

Modeling and prediction of water quality are very important for the protection of the environment. Developing a model by using advanced artificial intelligence algorithms can be used to measure the future water quality. In this proposed methodology, the advanced artificial intelligence algorithms, namely, NARNET and LSTM models were used to predict the WQI. Moreover, machine learning algorithms such as SVM, KNN, and Naive Bayes were used to classify the WQI data. The proposed models were evaluated and examined by some statistical parameters. For the WQI prediction, the result has revealed that the performance of the NARNET model is slightly better than the LSTM model based on the obtained value. However, the SVM algorithm has achieved the highest accuracy of the prediction of the WQC as compared with KNN and Naive Bayes algorithms. After examining the robustness and efficiency of the proposed model for predicting the WQI, in future work, the developed models will be implemented to predict the water quality in Saudi Arabia for different types of water.

## **5. FLOW CHART**

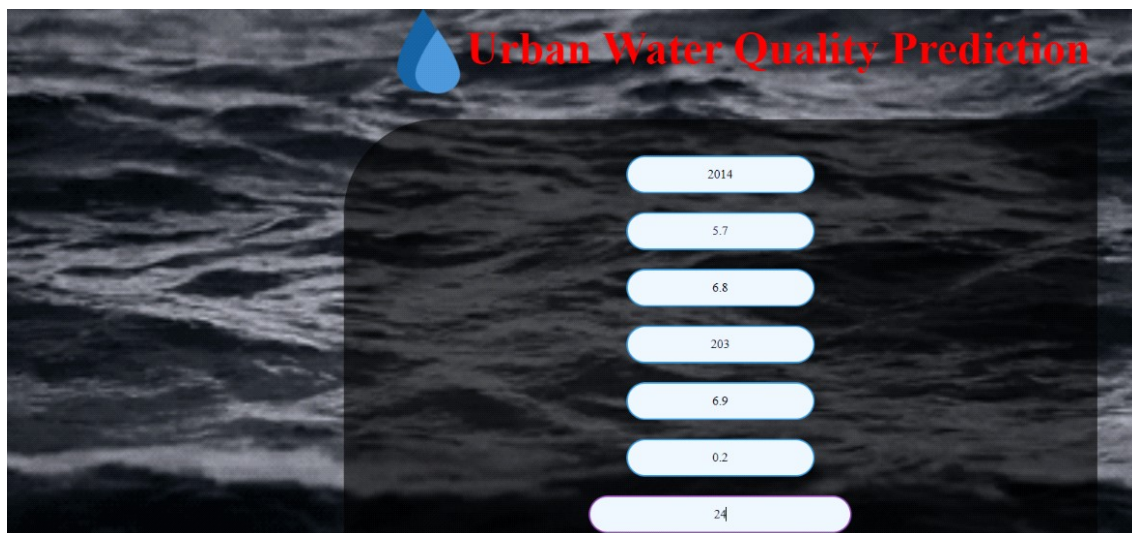
- Data Collection.
  - Collect the dataset or Create the dataset
- Data Preprocessing.
  - Import the Libraries.
  - Importing the dataset.
  - Checking for Null Values.
  - Data Visualization.
  - Taking care of Missing Data.
  - Label encoding.
  - One Hot Encoding.
  - Feature Scaling.
  - Splitting Data into Train and Test.
- Model Building
  - Training and testing the model
  - Evaluation of Model
- Application Building
  - Create an HTML file
  - Build a Python Code



## 6. RESULT

For validating the developed model, the dataset has been divided into 70% training and 30% testing subsets. While the ANN and LSTM models were used to predict the WQI, the SVM, KNN, and Naive Bayes were utilized for the water quality classification prediction.

A NARNET model, with 12 hidden layers, showed a good performance to predict the WQI values. As presented earlier, it has the following characteristics: 1 : 8 number of delays and 12 number of epochs. However, the developed LSTM model has a total number of 200 hidden layers, 150 maximum number of epochs, and delay. The proposed models were evaluated and examined by some statistical parameters. For the WQI prediction, the result has revealed that the performance of the NARNET model is slightly better than the LSTM model based on the obtained R value. However, the SVM algorithm has achieved the highest accuracy of the prediction of the WQC as compared with KNN and Naive Bayes algorithms. After examining the robustness and efficiency of the proposed model for predicting the WQI, in future work, the developed models will be implemented to predict the water quality in Saudi Arabia for different types of water.



Here we can see the final output of the project in which it will analyse and gives the predicted value of the water.

## **7. ADVANTAGES AND DISADVANTAGES**

### *Advantages*

- The predicted values are accurate.
- It will analyse the given water sample.
- It is the advanced technology used for prediction.
- We can also get the graphical representation by using the data visualization.

### *Disadvantages*

- Accuracy should be maintained to get exact values.
- Difficult to built the model.

## **8. APPLICATIONS**

- Urban water prediction for drinking, household purposes etc..
- Industry and commerce.
- For power generation.

## **9. CONCLUSION**

The performance of artificial intelligence techniques including GMDH, SVM and ANN were evaluated to predict the water quality components of Tireh River (Iran). To this end most dataset related well-known components, such as pH,  $\text{SO}_4$ , Na, Ca, Cl, Mg,  $\text{HCO}_3$  etc., were collected. Results indicated that the applied models have suitable performance for predicting water quality components, however, the best performance was related to the SVM. Reviewing the structure of SVM showed that the best accuracy was related to the RBD as kernel function. Results of ANN indicated that its accuracy is acceptable for practical purposes. The best performance of tested transfer function was related to tansig. The lowest accuracy of models was related to GMDH. The DDR index of results of applied models shows that all three models slightly over-estimate. Comparison of the performance of GMDH, SVM and ANN according to DDR shows that the data dispersion of SVM was less than the others. Although the accuracy of model GMDH is less than that of model SVM, their DDR indices are close together. Furthermore, comparison of the performance of applied models indicated that the outcomes of GMDH and SVM models were more reliable in comparison with ANN.

## **10. FUTURE SCOPE**

The paper presents the economical solution to avoid contamination of water in residential overhead tanks. The quality of water is monitored using IoT devices and the future prediction of water contamination is achieved using machine learning algorithms. The proposed system consists of multi sensors connected to NodeMCU to collect the water parameters. And the alert message is sent to the user before the water gets contaminated. The system helps to save the water from contamination and is also cost effective. The future scope for this project is to detect the diseases caused by different parameters and finding the appropriate solution for to clean the tank. Also biosensors can be used to detect the microbacterias for better quality of water.

## **11. REFERENCES**

- [1] P. Zeilhofer, L. V. A. C. Zeilhofer, E. L. Hardoim, Z. M. . Lima, and C. S. Oliveira, "GIS applications for mapping and spatial modeling of urban-use water quality: a case study in District of Cuiabá, Mato Grosso, Brazil," *Cadernos de Saúde Pública*, vol. 23, no. 4, pp. 875–884, 2007.
- [2] M. A. Kahlown, M. A. Tahir, and H. Rasheed, National Water Quality Monitoring Programme, Fifth Monitoring Report (2005–2006), Pakistan Council of Research in Water Resources Islamabad, Islamabad, Pakistan, 2007, <http://www.pcrwr.gov.pk/Publications/Water%20Quality%20Reports/Water%20Quality%20Monitoring%20Report%202005-06.pdf>.
- [3] UN water, "Clean water for a healthy world," Development, 2010, <https://www.undp.org/content/undp/en/home/presscenter/articles/2010/03/22/clean-water-for-a-healthyworld.html>.
- [4] Nikhil Kumar Koditala ; Purnendu Shekar Pandey, "Water Quality Monitoring System Using IoT and Machine Learning", IEEE paper, August 2018.
- [5] Najah, A. El-Shafie, O. A. Karim, Amr H. El-Shafie —Application of artificial neural networks for water quality prediction, May 2013, Volume 22, pp 187–201.
- [6] K. Farrell-Poe, W. Payne, and R. Emanuel, *Water Quality & Monitoring*, University of Arizona Repository, 2000, <http://hdl.handle.net/10150/146901>.

- [7] T. Taskaya-Temizel and M. C. Casey, "A comparative study of autoregressive neural network hybrids," *Neural Networks*, vol. 18, no. 5–6, pp. 781–789, 2005.
- [8] C. N. Babu and B. E. Reddy, "A moving-average filter based hybrid ARIMA-ANN model for forecasting time series data," *Applied Soft Computing*, vol. 23, pp. 27–38, 2014.
- [9] X. Zhang, N. Hu, Z. Cheng, and H. Zhong, "Vibration data recovery based on compressed sensing," *Acta Physica Sinica*, vol. 63, no. 20, pp. 119–128, 2014.
- [10] M. M. S. Cabral Pinto, C. M. Ordens, M. T. Condesso de Melo et al., "An interdisciplinary approach to evaluate human health risks due to long-term exposure to contaminated groundwater near a chemical complex," *Exposure and Health*, vol. 12, no. 2, pp. 199–214, 2020.
- [11] M. M. S. Cabral Pinto, A. P. Marinho-Reis, A. Almeida et al., "Human predisposition to cognitive impairment and its relation with environmental exposure to potentially toxic elements," *Environmental Geochemistry and Health*, vol. 40, no. 5, pp. 1767–1784, 2018.
- [12] Y. C. Lai, C. P. Yang, C. Y. Hsieh, C. Y. Wu, and C. M. Kao, "Evaluation of non-point source pollution and river water quality using a multimedia two-model system," *Journal of Hydrology*, vol. 409, no. 3–4, pp. 583–595, 2011.
- [13] J. Huang, N. Liu, M. Wang, and K. Yan, "Application WASP model on validation of reservoir-drinking water source protection areas delineation," in *2010 3rd International Conference on Biomedical Engineering and Informatics*, pp. 3031–3035, Yantai, China, October 2010.
- [14] I. R. Warren and H. K. Bach, "MIKE 21: a modelling system for estuaries, coastal waters and seas," *Environmental Software*, vol. 7, no. 4, pp. 229–240, 1992.

## 12. APPENDIX

```
import numpy as np
from flask import Flask,render_template,request
import pickle

app = Flask(__name__)
model = pickle.load(open('wqi.pkl','rb'))

@app.route('/')
def home() :
    return render_template("web.html")

@app.route('/login',methods = ['POST'])
def login() :
    year = request.form["year"]
    do = request.form["do"]
    ph = request.form["ph"]
    co = request.form["co"]
    bod = request.form["bod"]
    na = request.form["na"]
    tc = request.form["tc"]
    total =
[[int(year),float(do),float(ph),float(co),float(bod),float(na),float(tc)]]
    y_pred = model.predict(total)
    print(y_pred)
    y_pred =y_pred[[0]]
    if(y_pred >= 95 and y_pred <= 100) :
        return render_template("web.html",showcase = 'Excellent,The predicted
value is '+ str(y_pred))
    elif(y_pred >= 89 and y_pred <= 94) :
        return render_template("web.html",showcase = 'Very good,The predicted
value is '+str(y_pred))
    elif(y_pred >= 80 and y_pred <= 88) :
        return render_template("web.html",showcase = 'Good,The predicted value
is'+str(y_pred))
    elif(y_pred >= 65 and y_pred <= 79) :
        return render_template("web.html",showcase = 'Fair,The predicted value
is '+str(y_pred))
    elif(y_pred >= 45 and y_pred <= 64) :
        return render_template("web.html",showcase = 'Marginal,The predicted
value is '+str(y_pred))
    else :
        return render_template("web.html",showcase = 'Poor,The predicted value
is '+str(y_pred))

if __name__ == '__main__' :
    app.run(debug = True,port=5000)
import requests
```

# NOTE: you must manually set API\_KEY below using information retrieved from your IBM Cloud account.

```
API_KEY = "9i1jB450-KLd74q0v2Cy8mWxfRfkAwfpu0xW08jCFHIW"
```

```
token_response = requests.post('https://iam.cloud.ibm.com/identity/token',  
data={"apikey": API_KEY, "grant_type": 'urn:ibm:params:oauth:grant-type:apikey'})
```

```
mltoken = token_response.json()["access_token"]
```

```
header = {'Content-Type': 'application/json', 'Authorization': 'Bearer ' +  
mltoken}
```

# NOTE: manually define and pass the array(s) of values to be scored in the next line

```
payload_scoring = {"input_data": [{"fields": [['do', 'ph', 'co', 'bod', 'na',  
'tc', 'year']], "values": [[6.7,7.5,203.0,6.940049,0.1,27.0,2014]]}]}
```

```
response_scoring = requests.post('https://us-  
south.ml.cloud.ibm.com/ml/v4/deployments/574b29c7-1949-4f2c-bdbf-  
9380f5094ab6/predictions?version=2022-03-05', json=payload_scoring,  
headers={'Authorization': 'Bearer ' + mltoken})
```

```
print("Scoring response")
```

```
print(response_scoring.json())
```

```
pred=response_scoring.json()
```

```
y_pred=pred['predictions'][0]['values'][0][0]
```

```
print(y_pred)
```

```
import numpy as np
```

```
from flask import Flask,render_template,request
```

```
import pickle
```

```
import requests
```

# NOTE: you must manually set API\_KEY below using information retrieved from your IBM Cloud account.

```
API_KEY = "9i1jB450-KLd74q0v2Cy8mWxfRfkAwfpu0xW08jCFHIW"
```

```
token_response = requests.post('https://iam.cloud.ibm.com/identity/token',  
data={"apikey": API_KEY, "grant_type": 'urn:ibm:params:oauth:grant-type:apikey'})
```

```
mltoken = token_response.json()["access_token"]
```

```
header = {'Content-Type': 'application/json', 'Authorization': 'Bearer ' +  
mltoken}
```

```
app = Flask(__name__)
```

```
@app.route('/')  
def home() :
```

```
    return render_template("web.html")
```

```
@app.route('/login',methods = ['POST'])  
def login() :
```

```
    year = request.form["year"]
```

```
    do = request.form["do"]
```

```
    ph = request.form["ph"]
```

```

co = request.form["co"]
bod = request.form["bod"]
na = request.form["na"]
tc = request.form["tc"]
total =
[[int(year),float(do),float(ph),float(co),float(bod),float(na),float(tc)]]
#y_pred = model.predict(total)
payload_scoring = {"input_data": [{"fields": [['do', 'ph', 'co', 'bod',
'na', 'tc', 'year']], "values": total}]}
response_scoring = requests.post('https://us-
south.ml.cloud.ibm.com/ml/v4/deployments/574b29c7-1949-4f2c-bdbf-
9380f5094ab6/predictions?version=2022-03-05', json=payload_scoring,
headers={'Authorization': 'Bearer ' + mltoken})
print("Scoring response")
pred=response_scoring.json()
y_pred=pred['predictions'][0]['values'][0][0]
print(y_pred)
#y_pred =y_pred[[0]]
if(y_pred >= 95 and y_pred <= 100) :
    return render_template("web.html",showcase = 'Excellent,The predicted
value is '+ str(y_pred))
elif(y_pred >= 89 and y_pred <= 94) :
    return render_template("web.html",showcase = 'Very good,The predicted
value is '+str(y_pred))
elif(y_pred >= 80 and y_pred <= 88) :
    return render_template("web.html",showcase = 'Good,The predicted value
is '+str(y_pred))
elif(y_pred >= 65 and y_pred <= 79) :
    return render_template("web.html",showcase = 'Fair,The predicted value
is '+str(y_pred))
elif(y_pred >= 45 and y_pred <= 64) :
    return render_template("web.html",showcase = 'Marginal,The predicted
value is '+str(y_pred))
else :
    return render_template("web.html",showcase = 'Poor,The predicted value
is '+str(y_pred))

if __name__ == '__main__' :
    app.run(debug = True,port=5000)

```