Linear Regression aka Least Squares

Bias- inability to capture the true relationship. Straight line to model curve

Variance and Overfitting. Complex models fit the data with complex functions(Squiggly line) hence during testing the variability will be high

Ideal is low bias and low variability

Squiggly line fits the training set really well bit not the testing set and is an example for overfit

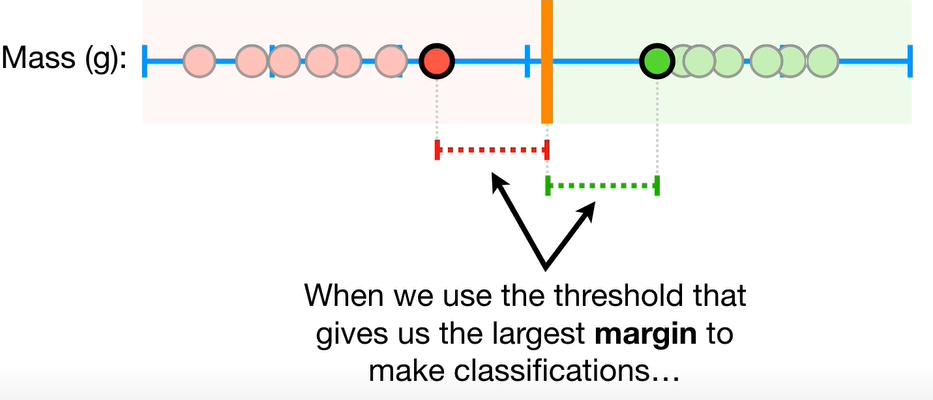
How to find the sweet spot between a line and a complex squiggly?

-Regularization, Boosting and Bagging

Cross validation is testing a particular ML model by dividing the dataset into subsections and using only one of it to test and the rest to train. And the process is repeated with all sections/ blocks. This is how different models are compared based on the testing results. The number by which the dataset is subdivided is called the fold. If subdivided by 10, it’s called 10 fold cross validation

SVM

Margin: Distance between the threshold and the data point that is closer to the threshold



Largest margin for classifications: Maximal Margin Classifier (super sensitive to outlier in training data)

Bias/Variance Tradeoff that allows slight misclassifications

Higher bias, low variance

Support vector classifiers allow soft margins and when data is difficult to segregate, Support Vector Machines are used which adds another dimension to the data using a kernel. Transform data using Kernel Function

Radial Basis Function Kernel (helps to find the SVC )

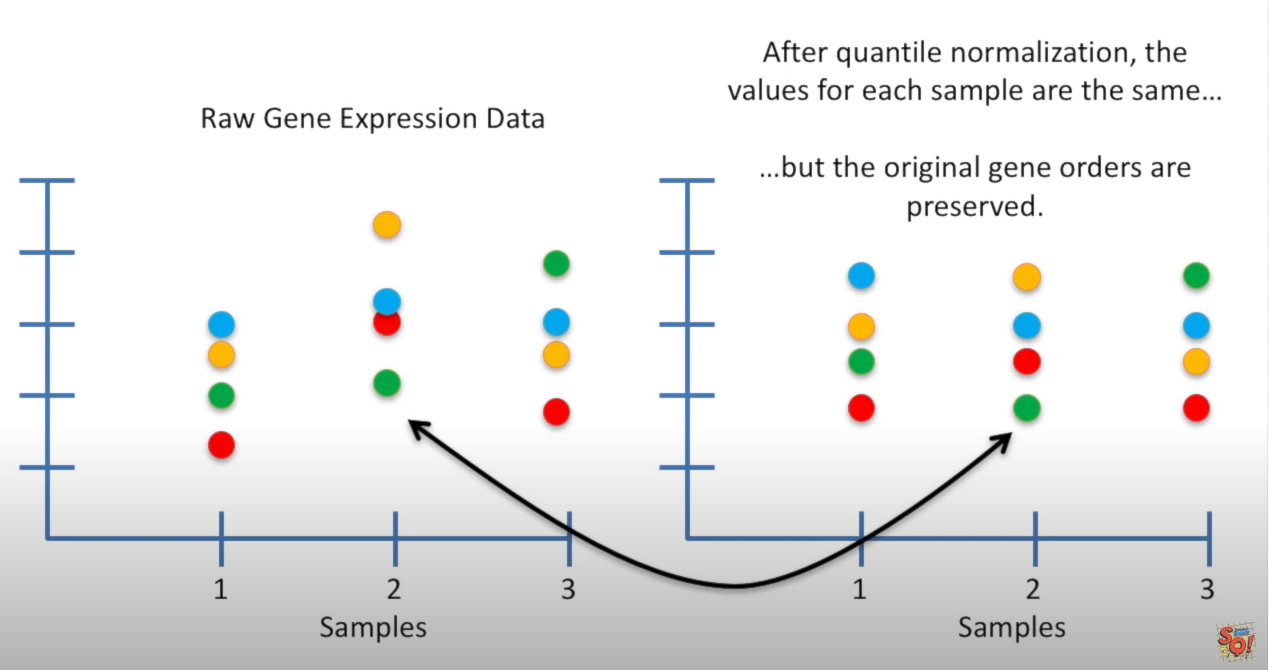
Allowing misclassifications the margin is called soft margin

Correlation vs Causation, Spurious Correlation

Autocorrelation in time series data: Correlation doesn’t mean causality  
Causality tests

[**https://www.kdnuggets.com/2020/03/machine-learning-time-series-forecasting-sequel.html**](https://www.kdnuggets.com/2020/03/machine-learning-time-series-forecasting-sequel.html)

**Normalized Dataset when different samples from experiments have different mean. Normalization brings the samples in the same levels based on the mean of all the samples. Cancels out the effect of mean difference. Also called standardization**

****

**Standardising the values to the range of 0-1 for example**

**R squared is a percentage of variation of the data explained by the relationship. Higher the better**

**RMSE the lower the better**

**Forecasting is the process of making predictions of the future based on past and present data and most commonly by analysis of trends.**

**Univariate and Multivariate time series**

**According to** [**NASA**](https://earthdata.nasa.gov/user-resources/remote-sensors)**, a spectral radiometer *is* a multispectral sensor.**

**Spectroradiometer—A radiometer that measures the intensity of radiation in multiple wavelength bands (i.e., multispectral). Many times the bands are of high-spectral resolution, designed for remotely sensing specific geophysical parameters**

**Perhaps you're thinking of a spectrometer, which is slightly different, but similar**

**Spectrometer—A device that is designed to detect, measure, and analyze the spectral content of incident electromagnetic radiation. Conventional imaging spectrometers use gratings or prisms to disperse the radiation for spectral discrimination.**

**For the sake of being thorough,**

**Radiometer—An instrument that quantitatively measures the intensity of electromagnetic radiation in some bands within the spectrum. Usually, a radiometer is further identified by the portion of the spectrum it covers; for example, visible, infrared, or microwave.**

**And**

**Hyperspectral radiometer—An advanced multispectral sensor that detects hundreds of very narrow spectral bands throughout the visible, near-infrared, and mid-infrared portions of the electromagnetic spectrum. This sensor’s very high spectral resolution facilitates fine discrimination between different targets based on their spectral response in each of the narrow bands.**

**#NOT DONE IN REAL-TIME**

**#LIMITED DATASETS**

**#EXPENSIVE**

**#Lot of input parameters**

**#data not from real time sensors  
#Historic/Time Series data are geographically specific and specific to water body. Rapid deterioration of environment changes the relationship (find this outR) (vulnerable paper check that . it says all ML are vulnerable)**

**Aim**

**Technology/Algorithm**

**Result**

**Gap:**

**Parameters**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*The Trophic State Index (TSI) is a classification system designed to rate water bodies based on the amount of biological activity they sustain.[1] Although the term "trophic index" is commonly applied to lakes, any surface water body may be indexed.*

*The TSI of a water body is rated on a scale from zero to one hundred.[1] Under the TSI scale, water bodies may be defined as:[1]*

*oligotrophic (TSI 0–40, having the least amount of biological productivity, "good" water quality);*

*mesoeutrophic (TSI 40–60, having a moderate level of biological activity, "fair" water quality); or*

*eutrophic to hypereutrophic (TSI 60–100, having the highest amount of biological activity, "poor" water quality).*

**Eﬃcient Water Quality Prediction Using Supervised Machine Learning  
Ahmed et al. (2019) (1 citaiton)**

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient Water Quality Prediction Using Supervised Machine Learning. *Water*, *11*(11), 2210.

**(Drinking water)**

**#Methodology**

**Aim:** Estimate Water Quality Index using Supervised Machine Learning

**Technology/Algorithm:** Gradient Boosting & Polynomial Regression for WQI (Water Quality Index)and Multi-layer Perceptron for WQC (Water Quality Class)

**Result:** Predicts with good accuracy and is suitable for real-time water quality detection using a few easily obtainable water quality parameter

**Gap:** Only for drinking water, Based on lab samples and only 663 samples from 12 different sources

**Parameters:** Temperature, Turbidity, pH & TDS  
 **Algal Bloom Prediction Using Extreme Learning Machine Models at Artiﬁcial Weirs in the Nakdong River, Korea (7 citations)  
Yi et al. (2018)  
(River)**

**#Body**

Yi, H. S., Park, S., An, K. G., & Kwak, K. C. (2018). Algal bloom prediction using extreme learning machine models at artificial weirs in the Nakdong River, Korea. International journal of environmental research and public health, 15(10), 2078.

**Aim:** Chlorophyll A concentration was predicted as it is a primary indicator of algal bloom

**Technology/Algorithm:** Extreme Learning Model (ELM)

**Result:** ELM performed well with data collected from upstream weirs with higher and lower RMSE

**Gap:** Regulated river with artificial weirs, trained model based on data collected over 3 years not with their own sensors

**Parameters:** Air temperature, rainfall, solar radiation, total nitrogen, total phosphorus, N/P ratio and chlorophyll-a concentration

**An optimization of artiﬁcial neural network model for predicting chlorophyll dynamics Tian et al. (2017)(18 citations)**

**Tian, W., Liao, Z., & Zhang, J. (2017). An optimization of artificial neural network model for predicting chlorophyll dynamics. *Ecological Modelling*, *364*, 42-52.**

**(estuary reservoir)**

**#Inductive and deductive   
#Body**

**Algal Blooms affect the quality of drinking water  
Detection of Algal Bloom requires monitoring routinely or reactively. The frequency of the WQ has to be high to detect the onset of bloom. Difficult due to:**

1. **Expense of field monitoring**
2. **Staff and Resource availability**
3. **Field safety**
4. **Significant time delay between collection, measurement, reporting & notifying public**

**Forecasting models are two types including deductive and inductive where deductive require in depth information of physical, chemical and biological processes. The accuracy of prediction is constrained by lack of knowledge of the complex algae growth mechanism**

**Inductive models provide a holistic view of the system by extracting patterns from empirical data using statistical correlation and machine learning. ANN falls under inductive models and can work with the complexities in natural processes like chlorophyll dynamics.**

**ANN used to predict chl-a for the last 20 years as it can give an early warning to algal bloom so that it can be dealt with proactively.**

**ANN is problem dependent as it depends on the climate, hydrology, water quality etc**

**Chlorophyll value is non stationary time series process and is influences by complex environmental factors which makes forecasting difficult**

**Issue of autocorrelation. Time delay was observed for training and prediction**

**(Problems with time series)**

**#Methodology Data was normalized.  
Model performance evaluated by MSE, R, Bias feature & Accuracy and lastly Significance P (T-test)**

**Aim:** Chlorophyll dynamics (Chl-a) was predicted using optimized ANN that predicted the change in chlorophyll value instead of the base value. The prediction of chlorophyll dynamics provides an early prediction of algal bloom so that it can be prevented

**Technology/Algorithm:** Artificial Neural Network (ANN) with backpropagation algorithm and Early Stopping Technique (EST) to improve generalization ability and avoid overfitting

**Result:** Optimized ANN performed better than traditional ANN as change in chlorophyll is more sensitive to input data

**Gap:** 1152 samples collected between 5-9th of June and 24th-26th of July 2015 in 10 min intervals

**Parameters:** Temperature, pH, Electrical Conductivity, Oxidation Reduction Potential (ORP), turbidity, DO and Chlorophyll-A (Inputs)

Chl-a change (Output)

**Comparison of Machine Learning Algorithms for Retrieval of Water Quality Indicators in Case-II Waters: A Case Study of Hong Kong (6 citations)  
Hafeez et al. (2019)**

**Hafeez, S., Wong, M. S., Ho, H. C., Nazeer, M., Nichol, J., Abbas, S., ... & Pun, L. (2019). Comparison of machine learning algorithms for retrieval of water quality indicators in case-II waters: a case study of Hong Kong. *Remote sensing*, *11*(6), 617.**

**(Subtropical Coastal Water)**

**#methodology #validation- Body**

**#retrieval**

**Aim :** Tested four ML algorithms to retrieve concentration of water quality parameters of Chl-a, Suspended Solids (SS) and Turbidity using input data from handheld spectroradiometer (reflectance ) and satellite data

**Technology/Algorithm:** ANN

**Result:** ANN outperformed other MLs (Support Vector Regression (SVR), Random Forest (RF) & Cubist Regression Trees(CB)) for both the data input (reflectance and satellite)

**Gap:** Relies of Satellites and Multispectral Radiometer(MSR). Satellite needs clear weather, MSR is likely to be expensive

**Parameters:** (Not input but retrieved) Chl-a, Suspended Solids (SS) and Turbidity

**Deep-Learning-Based Approach for Prediction of Algal Blooms**

**Zhang et al. (2016) (7 citations)**

**Zhang, F., Wang, Y., Cao, M., Sun, X., Du, Z., Liu, R., & Ye, X. (2016). Deep-learning-based approach for prediction of algal blooms. *Sustainability*, *8*(10), 1060.**

**(Coastal Water, sea, china)**

**Aim:** Deep Learning Model that finds out the relationship between phytoplankton cell density (a direct predictor of algal blooms) and predicts the density of phytoplankton when presented with water quality parameters

**Technology/Algorithm:** Multilayer Deep Belief Networks (DBN) architecture composed of multiple restricted Boltzmann machine (RBM)

**Result:** high accuracy in predicting algal blooms and it depends on meteorological, hydrological, marine biological and chemical conditions which are non linear

**Gap:** High computation power needed to deal with datasets, obtained from 20 monitoring stations, biweekly data from 2008-2012. Not for Lakes

**Parameters:** sea surfaceTemperature, Salinity, pH, Chlorophyll-a, chemical oxygen demand, dissolved oxygen (DO), phosphate (PO4\_P), acid nitrate (NO2\_N), nitrate (NO3\_N), ammonia nitrogen (NH4\_N), and silicate

**Estimation of high frequency nutrient concentrations from water quality surrogates using machine learning methods. Castrillo & García (2020) (No cit available)**

**Castrillo, M., & García, Á. L. (2020). Estimation of high frequency nutrient concentrations from water quality surrogates using machine learning methods. *Water Research*, 115490.**

**(Catchment/ River)  
#soft sensor programming  
#Body   
Continuous WQ monitoring is imperative to support water management. Good water quality (ecological and chemical) are requred to ensure good public health, healthy ecosystems, biodiversity and protection of water supply. One of the problems is all parameters can not be measured in situ.**

**Monitoring of nutrients using grab sampling has a low sampling frequency because of resources required, cost, difficulty and delays and hence online, reliable and cost-effective are in demand.**

**Commercially available sensor technologies include ion-selective electrodes (ISE) which are cost effective but are generally susceptible to drifting, inaccuracies and interferences. Wet chemical and optical sensors are precise but are very expensive and requires routine maintenance.**

**Inferring target parmeter with the available data by using a software model is called soft sensor**

**\*Urban areas tend to be more comples for prediction and linear relationships are not suitable.**

**\*Non linear machine learning models have advantages over linear regression models**

**\*choosing irrelevant data can lead to overfitting: Stepwise Forward Subset was executed to assess the performance indiviual input parameter on the predicted output**

**#methodology**

**Preprocessing:**

**Data Cleaning**

**#important Pearson’s Correlation coefficient for Parameter Correlation**

**\*Seasonal behaviour also correlates with nutrient concentration and Temp**

**\*Covariance between two variables has to be taken into consideration in soft sensor design**

**#limitation- absence of data from main sensor due to calibration, maintenance, drifting, failure etc.**

**#validation  
Open dataset:** [**https://catalogue.ceh.ac.uk/eidc/documents?facet=topic%7C0%2FWater+quality%2F**](https://catalogue.ceh.ac.uk/eidc/documents?facet=topic%7C0%2FWater+quality%2F)

**Historic data from 2009 to 2012 historic data**

**Aim:** Commonly measured in-situ parameters are used (as surrogates/substitute) to predict nutrient concentration in rural and urban catchment. This study assesses the feasibility of non-linear ML models to linear model

**Technology/Algorithm:** Random Forest

**Result:** Reduction in RMSE by almost 60% compared to linear modeling (Multiple Linear Regression(MLR)). RF has the ability to integrate more inputs to predict (using time series data which was seen in this study). Promising for soft-sensor  
TRP & NO3-N has a strong correlation with EC

Turb is used to predict TRP  
pH is used to predict NO3-N

**Gap:** Reliant on specific variable selection, compared against Multiple linear regression. Accuracy not that high

**Parameters: Input: EC, FLow, Temp, Turb Output: TRP, TP & NO3-N, NH4-N (nutrient) and based on river (flow in an important vairable)**

EC, Turbidity, Temperature, DO saturation concentration (DOsat), pH, chlorophyll concentration and Flow rate used to predict Total Reactive Phosphorus (TRP) and nitrogen as nitrate (NO3-N) or nitrogen as Ammonia (NH4-N) and Total Phosphorus(TP)

**Hyperspectral Data and Machine Learning for Estimating CDOM, Chlorophyll a, Diatoms, Green Algae and Turbidity. Keller et al. (2018) (15 citations)  
Keller, S., Maier, P. M., Riese, F. M., Norra, S., Holbach, A., Börsig, N., ... & Hinz, S. (2018). Hyperspectral data and machine learning for estimating CDOM, chlorophyll a, diatoms, green algae and turbidity. *International journal of environmental research and public health*, *15*(9), 1881.**

**(River)  
#methodology, preprocessing**

**Aim:** Five water quality parameters estimation from were evaluated using hyperspectral data in this study

**Technology/Algorithm:** Sensors used are Fluorometer PhycoSens, Biofish multi-sensor system and Visible and near infrared (VNIR) Cubert Hyperspectral sensor were used. Principal Component Analysis was used for preprocessing that reduces dimensionality of hyperspectral data. Linear regression (least squares), partial least squares (PLS), random forest (RF), extremely randomized trees (ET), adaptive boosting (AdaBoost), gradient boosting (GB), k-nearest neighbors (k-NN), support vector machine (SVM), artificial neural networks (ANN) and self-organizing maps (SOM) were the machine learning algorithm studied

**Result:** ANN, SVM and ET performed better than other algorithms with R^2 values ranging from 89.9% to 94.6%

**Gap:** Hyperspectral sensors are costly

**Parameters:** Colored Dissolved Organic Matter (CDOM), Chlorophyll-a, green algae, diatoms and turbidity

**Prediction of Algal Chlorophyll-a and Water Clarity in Monsoon-Region Reservoir Using Machine Learning Approaches. Mamun et al. (2020) (No cit available)**

Mamun, M., Kim, J. J., Alam, M. A., & An, K. G. (2020). Prediction of Algal Chlorophyll-a and Water Clarity in Monsoon-Region Reservoir Using Machine Learning Approaches. *Water*, *12*(1), 30.

**(Drinking water, reservoir located upstream of a river)**

**#Validation  
time series**

**Aim:** Long term Chlorophyll-A and Transparency in Secchi depth was predicted using data-driven machine learning models in different zones (riverine, transitional and lacustrine zone) and seasons to determine the trophic state. The models verified using cross-validation process

**Technology/Algorithm:** Multiple Linear Regression (MLR), Support Vector Machine (SVM) and Artificial Neural Network (ANN)

**Result:** SVM performed better than MLR and ANN with lowest Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE)

**Gap:** YSI Sonde (expensive) to measure, Too many measured parameters not in-situ

**Parameters:** Electrical Conductivity (EC), Dissolved Oxygen (DO), Water Temperature (WT), Total Nitrogen (TN), biological oxygen demand (BOD), chemical oxygen demand (COD), Total Phosphorus (TP)

**Using Machine-Learning Algorithms for Eutrophication Modeling: Case Study of Mar Menor Lagoon (Spain). Jimeno-Sáez et al. (2020) (No cit available)**

Jimeno-Sáez, P., Senent-Aparicio, J., Cecilia, J. M., & Pérez-Sánchez, J. (2020). Using Machine-Learning Algorithms for Eutrophication Modeling: Case Study of Mar Menor Lagoon (Spain). *International Journal of Environmental Research and Public Health*, *17*(4), 1189.  
**(Lagoon)**

**#Validation  
#data** [**http://www.canalmarmenor.es/web/canalmarmenor/parametros**](http://www.canalmarmenor.es/web/canalmarmenor/parametros)

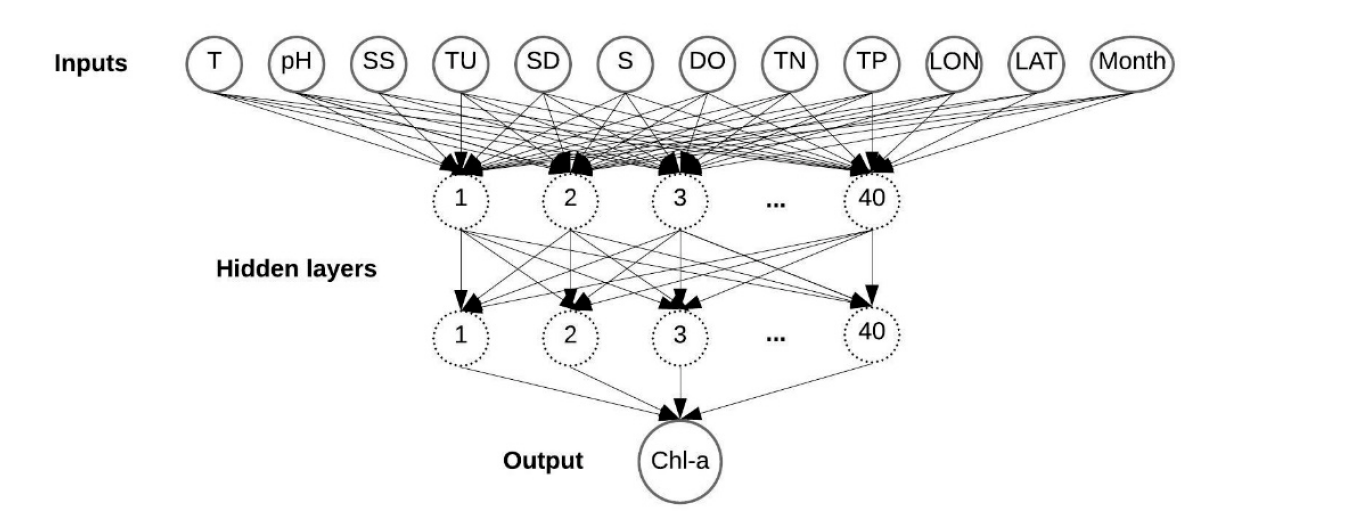
**Aim:** Multilayer Neural Networks (MLNNs) and Support Vector Regressions (SVRs) were evaluated to predict Chl-a using 9 input parameters

Technology/Algorithm: Multilayer Neural Networks (MLNNs) and Support Vector Regressions (SVRs)

**Result:** SVR performed better than MLNNs, 0.7 (Cross validated coefficient of determination)

**Gap:** 126 data points, needs 9 input parameters

**Parameters:**  Chl-a, water temperature (T), pH, suspended solids (SS), turbidity (TU), Secchi Disk depth (SD), salinity (S), dissolve oxygen (DO), total nitrogen (TN) and total phosphorus (TP)



**Water eutrophication assessment relied on various machine learning techniques: A case study in the Englishmen Lake (Northern Spain) Nieto et al. (2019) (5 citations)**  
Nieto, P. G., García-Gonzalo, E., Fernández, J. A., & Muñiz, C. D. (2019). Water eutrophication assessment relied on various machine learning techniques: A case study in the Englishmen Lake (Northern Spain). *Ecological Modelling*, *404*, 91-102.

**(Lake)**

**#ML analysis details of algo**

**Aim:** Predicts algal bloom in lakes by predicting the Total Phosphorus

Technology/Algorithm: Support Vector Machines (SVM) with Artificial Bee Colony (ABC) algorithm which tunes the SVM

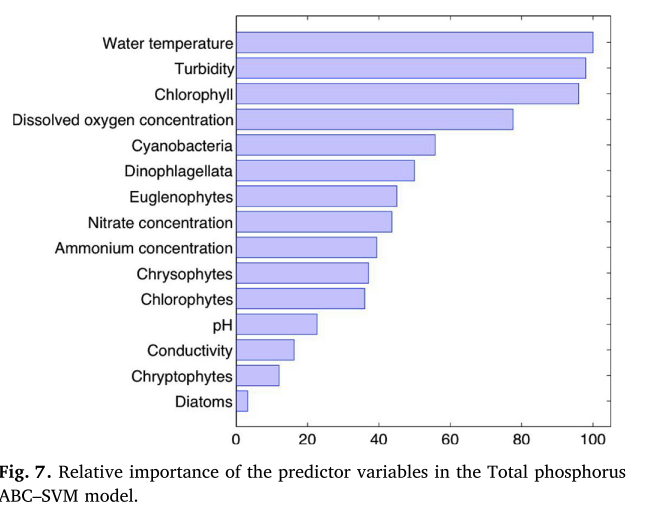
**Result:** Coefficient of determination is 0.92

**Gap:** Relies on measured in situ and lab data, depends on lot of parameters and expensive sensors and geographically limited

**Parameters: Biological parameters:** euglenophytes, Cyanobacteria, dinophlagellata, chlorophytes, diatoms, chrysophytes and cryptophytes

**• Physical–chemical parameters:** dissolved oxygen concentration ,turbidity, nitrate concentration, ammonium ion concentration, conductivity, water

temperature and pH.

****

**Using Artiﬁcial Neural Network Models for Eutrophication Prediction. Huo et al. (2013) (24 citations)**

Huo, S., He, Z., Su, J., Xi, B., & Zhu, C. (2013). Using artificial neural network models for eutrophication prediction. *Procedia Environmental Sciences*, *18*, 310-316.  
**(Lake)**

**#Standardization of Input parameters  
#parameter Correlation #check #parameter sensitivity analysis-> effect of input parameter on the output  
#month might affect Chl-1 growth. Turbidity Chl-A is correlated**

**#Chl-a is significantly correlated with month, SD, water temp, DO, pH, TN & TP  
BP-ANN model predictions are better correlated with observed Chl-a  
The proposed model can be used for Chl-a prediction**

**Can analyze the factors that control harmful algal bloom and provide fast predictions that can be used for forecasting**

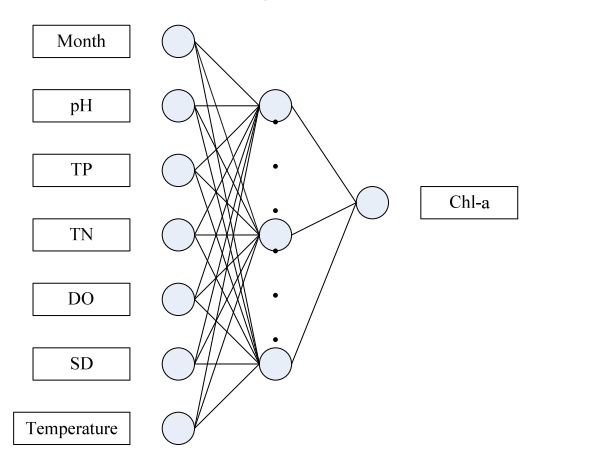
**Aim:** Several ML models were evaluated for predicting Total Nitrogen (TN), Secchi Depth (SD), Dissolved Oxygen (DO), Chl-a to determine Eutrophication.

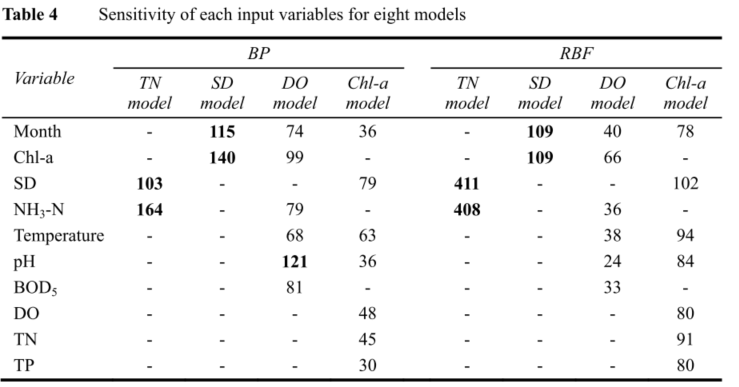
**Technology/Algorithm:** Back-Propagation(Feedforward ANN with Backprop algo) Neural Network and Radial Basis Function Neural Network

**Result:** BP-ANN outperformed RBF-ANN, correlation coefficient above 0.7 for BP-ANN

**Gap:** Limited Dataset

**Parameters:** Input: SD and NH3-N to predict TN, Month and Chl-a to predict SD, Month, pH, Biochemical Oxygen Demand (BOD5), NH3-N, Chl-a, Temperature to predict DO, Month, pH, TP, TN, SD, Temperature , DO to predict Chl-a

****

****

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_MODERATELY\_RELEVANT\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Design of River Water Quality Assessment and Prediction Algorithm**

**Cao et al. (2018) (no citation)  
Cao, S., Wang, S., & Zhang, Y. (2018, December). Design of River Water Quality Assessment and Prediction Algorithm. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 901-906). IEEE.**

**(River) #time series confirmed  
Aim:** A water quality assessment model was developed that analyzes water quality parameters and makes further predictions of the trend regarding water pollution. Two models developed. 1. mulitclass LS-SVC to evaluate water quality of each site and 2. Forecasting of water quality parameters

**Technology/Algorithm:** Mutation Genetic Algorithm is used with Particle Swarm Optimization (PSO) and Least Squares Support Vector machine (LS-SVM)/ Least squares Support Vector Classification (LS-SVC) based on PSO was used to tune the hyperparameters. LS-SVC was used to set up an assessment model for water quality to evalute the quality. Fuzzy Information Granulation method was combined with Least Square Support Vector Regression to predict the water quality changes upto 3 days in the future based on historic trends  
**Result:** This proved to be faster in training speed and more accurate in predicting than Back Propagation NN. PSO algorithm improved accuracy by 1.5%

**Gap:** Water Quality assessment model used 200 sample data & water quality prediction model used 160 groups of data. Limited data sets

**Parameters:** pH, DO, Chemical Oxygen Demand (COD), NH-3

**Predicting and analyzing water quality using Machine Learning: A comprehensive model.**

**Khan & See (2016) (7 citations) #time series confirmed  
Khan, Y., & See, C. S. (2016, April). Predicting and analyzing water quality using Machine Learning: A comprehensive model. In *2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT)* (pp. 1-6). IEEE.**

**(channel)  
Data source US Geological Survey’s (USGS) National Water Information System (NWIS)**

**Aim:** This paper analyzes and forecasts water quality parameters using ANN and time-series analysis to determine the concentration of the parameters. Predicts the future water quality trends of the particular region and the goal of the study is to predict values based on present values. Mean-Squared Error (MSE), Root Mean-Squared Error (RMSE) and Regression Analysis was used to evaluate the model.

**Technology/Algorithm:** Feedforward Neural Network with Non Linear Autoregressive (NAR) time series model used with Scaled Conjugate Gradient (SCG) training algorithm amd Log Sigmoid Activation function

**Result:** Predicted all four parameters from time series data fairly accurately. Chl- 0.873 R value

**Gap:** month wise data, limited datasets

**Parameters:** Chlorophyll, Specific Conductance, Dissolved Oxygen, Turbidity (FNU)

**A Supervised Learning Approach to Water Quality Parameter Prediction and Fault Detection. (No citation data available)**

**Joslyn & Lipor (2018)**

**Joslyn, K., & Lipor, J. (2018, December). A Supervised Learning Approach to Water Quality Parameter Prediction and Fault Detection. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 2511-2514). IEEE.**

**(River)#USGS #DATA #time series confirmed**

**Aim:** Water Quality Parameters (DO, cyanobacteria, turbidity) are predicted to automatically detect anomalies like sensor faults, fouling, decalibration etc using supervised machine learning. Technology/Algorithm: Support Vector Machine (SVM) & Gradient Boosting Algorithm (XGBoost) and Support Vector Regression (SVR)

**Result:** SVR used for DO, XGBoost for Cyanobacteria and SVR for turbidity

Showed good prediction power

**Gap: Not real time**

**Parameters:** dissolved oxygen (DO), pH balance, chlorophyll, temperature, speciﬁc conductivity, turbidity, cyanobacteria, nitrate, and ﬂuorescent dissolved oxygen matter (fDOM)as inputs

**\*Recommendations: Including temporal correlations in the input data for predicting future values and applying scaling factor that grows linearly to simulate a gradual sensor drift**

**\*additional notes: DO and Turbidity plays an important role in policy making for water body maintenance  
Persistent water quality monitoring is important for scientists and policy makers**

**Water quality prediction method based on LSTM neural network (18 Citations)**

**Wang et al. (2017)   
Wang, Y., Zhou, J., Chen, K., Wang, Y., & Liu, L. (2017, November). Water quality prediction method based on LSTM neural network. In *2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)* (pp. 1-5). IEEE.**

**(Lake)**

**Aim:** Paper proposes Long and Short Term Memory Neural Network (LSTM NN) to predict DO and total phosphorus (TP) based on historical data. The proposed model deals with time series data of DO collected monthly from 2000 to 2006. LSTM NN performance was compared with Back Propagation NN and Online Sequential Extreme Learning Machine (OS-ELM). It is a time series prediction problem and conventional Neural Networks are not suitable

**Technology/Algorithm:** Long and Short Term Memory Neural Network (LSTM NN)

**Result:** It was found that LSTM NN perform better than Back Propagation NN and Online Sequential Extreme Learning Machine (OS-ELM).

**Gap:** Long training cycle

**Parameters:** Input and Output both DO, TP, forecasts DO, TP

**Machine learning approaches for anomaly detection of water quality on a real-world data set. (6 Citations) Muharemi et al. (2019)**

**Muharemi, F., Logofătu, D., & Leon, F. (2019). Machine learning approaches for anomaly detection of water quality on a real-world data set. *Journal of Information and Telecommunication*, *3*(3), 294-307.**

**(Drinking water, water distribution system)**

**\*Imbalanced data- only with class data where one class is bigger in size than the other**

**#time series confirmed  
#ML using time series are vulnerable especially in the presence of highly imbalanced dataset**

**Real World dataset are noisy and highly imbalanced  
#Data Preparation process and evaluation of different ML model performance**

**Muharemi et al. (2019)**

**Aim:** Event detection (detect anomalies on water quality data) using time series data for water quality. It is challenging because of the time-series data. Takes in 9 water quality predictor variables and outputs event. The study also evaluates different machine learning models using F-score metric

**Technology/Algorithm** Logistic regression, linear discriminant analysis, support vector machines (SVM), artiﬁcial neural network (ANN), deep neural network (DNN), recurrent neural network (RNN) and long short-term memory (LSTM)

**Result:** 122,334samples used. Machine learning algorithms are generally vulnerable (Real world data are noisy and highly imbalanced which makes prediction difficult) but SVM, ANN and Logistic Regressions are less vulnerable.

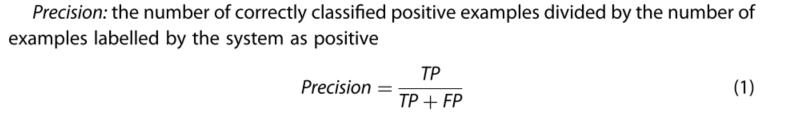
**Gap:** Finds the vulnerability but doesn’t offer a solution

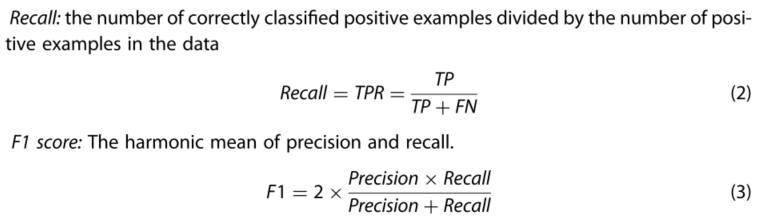
**Parameters:** Time,Temperature, Chlorine Dioxide, pH, Redox Potential, EC, Turbidity, flow rate, event change

Data Preparation process:

* 1. Data Cleaning
  2. Data Transformation
  3. Data Integration
  4. Noise Identification
  5. Imputing Missing Values

1. Feature Selection (Algorithms perform poorly because of correlated data. The study found Redox is positively correlated with pH while Chlorine dioxide is negatively correlated with temperature).





**Prediction of the five-day biochemical oxygen demand and chemical oxygen demand in natural streams using machine learning methods. Najafzadeh & Ghaemi (2019) (6 citations)  
Najafzadeh, M., & Ghaemi, A. (2019). Prediction of the five-day biochemical oxygen demand and chemical oxygen demand in natural streams using machine learning methods. *Environmental monitoring and assessment*, *191*(6), 380.**

**(Rivers)  
#benchmarking f test in comparing different ML models**

**Aim:** Multivariate Adaptive regression spline (MARS) and Least square-support vector machine (LS-SVM) were used to predict 5 day Biochemical Oxygen demand (BOD5) amd Chemical Oxygen Demand (COD). The selected algorithms were compared with ANN, Adaptive Neuro-Fuzzy inference system (ANFIS), Multiple nonlinear regression (MNLR) & multiple linear regression (MLR)

**Technology/Algorithm:** Multivariate Adaptive regression spline (MARS) and Least square-support vector machine (LS-SVM)

**Result:** LS-SVM with polynomial and RBF kernel functions were found to possess higher levels of accuracy in predicting BOD5 and COD

**Gap:** 200 samples, limited dataset

**Parameters:** EC, Sodium ion, Calcium ion, Magnesium ion, Orthophosphate (PO)

Nitrite, nitrate nitrogen, turbidity, and pH. (Input)

**Temporal and spatial variation of nutrients, suspended solids, and chlorophyll in Yeongsan watershed (9 citations). Mamun et al. (2018)  
Mamun, M., Lee, S. J., & An, K. G. (2018). Temporal and spatial variation of nutrients, suspended solids, and chlorophyll in Yeongsan watershed. *Journal of Asia-Pacific Biodiversity*, *11*(2), 206-216.**

**(Parameter Spearman’s correlation)  
(watershed)  
(Correlation analysis)**

This paper Carries out statistical analysis and correlation between parameters. It concludes that rainfall intensity is the main regulatory factor for nutrient concentration which is the determiner for algal growth. Total Phosphorus (TP) is the main nutrient in regulating algal Chlorophyll dynamics (CHL)

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_slightly\_relevant\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**A Gis-based system for assessing marine water quality around offshore platforms. Lu et al. (2014) (10 citations)**

**Lu, F., Chen, Z., & Liu, W. (2014). A Gis-based system for assessing marine water quality around offshore platforms. *Ocean & coastal management*, *102*, 294-306.**

**(Marine water- Sea/Bay)  
#Eutrophication Risk Level Calculation**

**Aim:** The study developed Eutrophication and heavy metal risk assessment/evaluation using fuzzy risk calculation and outputs contour and color coded maps in a Geographical Information System (GIS) based GUI

**Technology/Algorithm:** Fuzzy Set with evaluation criteria

**Result:** Contour maps depicted the water quality and it was found that water quality deterioration was mainly caused by high nutrient concentration

**Gap:** Not predicting, just classifying based on standards

**Parameters:** DO, COD, Phosphate, Inorganic Nitrogen, oil, Hg, Cu, Zn, Cd, Cr, As

**Hybrid soft computing approach for determining water quality indicator: Euphrates River. Li et al. (2019) (8 citations)**

Li, J., Abdulmohsin, H. A., Hasan, S. S., Kaiming, L., Al-Khateeb, B., Ghareb, M. I., & Mohammed, M. N. (2019). Hybrid soft computing approach for determining water quality indicator: Euphrates River. *Neural Computing and Applications*, *31*(3), 827-837.  
**(Rivers)  
#Sensitivity: importance of input WQ parameter to predict WQI**

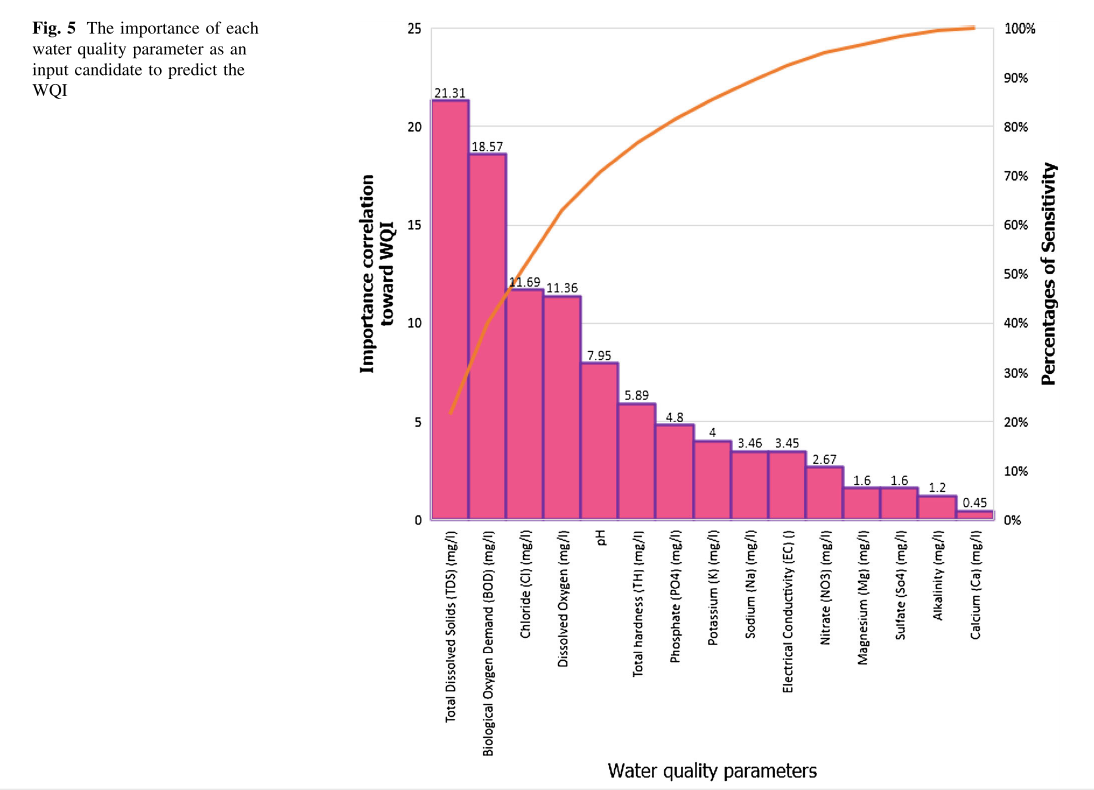
**Aim:** Water Quality Index was predicted using hybrid evolutionary model based on integrated support vector regression (SVR) with Firefly algorithm (FFA) and validated against other SVR based models

**Technology/Algorithm:** Hybrid SVR-FFA

**Result:** SVR-FFA performed better than other SVR models with R^2 being 0.90

**Gap:** Relies on a lot of data points

**Parameters:** TDS, BOD, Cl, Na, Mg, Ca, K, DO, EC, Alkalinity, pH, Phosphate, Nitrate, Total Hardness (TH), Sulfate



**Study of short-term water quality prediction model based on wavelet neural network (2013) (83 citations). Xu & Liu. (2013)  
Xu, L., & Liu, S. (2013). Study of short-term water quality prediction model based on wavelet neural network. *Mathematical and Computer Modelling*, *58*(3-4), 807-813.  
(freshwater pond for pearl breeding)  
Water quality heavily affected by hydrological and meteorological factors**

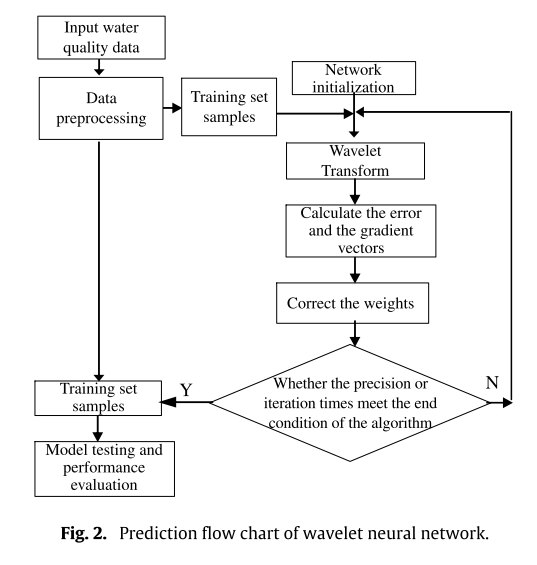
**Aim:** short-term prediction of DO in freshwater pond used for pearl breeding was carried out using Back propagation neural network combined with wavelet transform to make wavelet neural network and compared to Elman Neural Network

**Technology/Algorithm:** BP NN combined with Wavelet transform to make Wavelet Neural Network

**Result:** Mean absolute percentage error was reduced from 17.464% to 3.822%. Accuracy of prediction was greater than 90%

**Gap:** 144 data samples for training, collected over narrow time period (1 week) limited data. Data collected every 60 min

**Parameters:** DO, pH, Temperature, air humidity, wind speed and solar radiation levels

****

**Time series and machine learning to forecast the water quality from satellite data. Shehhi & Kaya (2020) (No citations data)**

**(Coastal Water)**

**Shehhi, M. R. A., & Kaya, A. (2020). Time series and machine learning to forecast the water quality from satellite data. *arXiv preprint arXiv:2003.11923*.**

**Aim:** Predicts Chl-a, Fluorescence line Height (FLH), sea surface temperature (SST) using time series ML model

**Technology/Algorithm:** Seasonal Auto Regressive Integrated Moving Average (SARIMA), regression and neural network

**Result:** Regression and NN are better at predicting Chl-a in all water types and SARIMA is better at predicting FLH and SST

**Gap:** Dusty and cloudy weather and arid and turbid regions

**Parameters:** satellite data input

**Using image processing technology for water quality monitoring system. Lai & Chiu (2011) (7 citations)**

**Lai, C. L., & Chiu, C. L. (2011, July). Using image processing technology for water quality monitoring system. In *2011 International Conference on Machine Learning and Cybernetics* (Vol. 4, pp. 1856-1861). IEEE.**

**(Aquarium)**

**Aim:** Water contamination was determined by observing fish gestures and behaviours using image processing and fuzzy inference technology

**Classifier for drinking water quality in real time. (8 citations)**

**Camejo et al. (2013)  
Camejo, J., Pacheco, O., & Guevara, M. (2013, January). Classifier for drinking water quality in real time. In *2013 International Conference on Computer Applications Technology (ICCAT)* (pp. 1-5). IEEE.**

**(Drinking water)**

**Aim:** Classified based in WQI Drinking water quality was assessed in this paper and was classified into two (good and medium) out of five classes (Excellent, good, medium, bad, very bad).

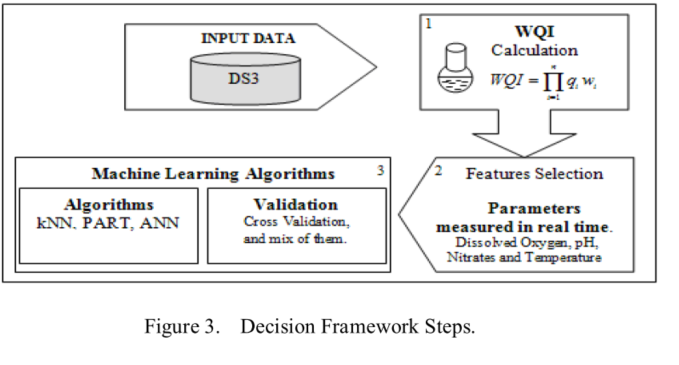
**Technology/Algorithm:** Partial Decision Trees (PART), Feed forward Artificial Neural Networks (ANN) and k-Nearest Neighbor (KNN)

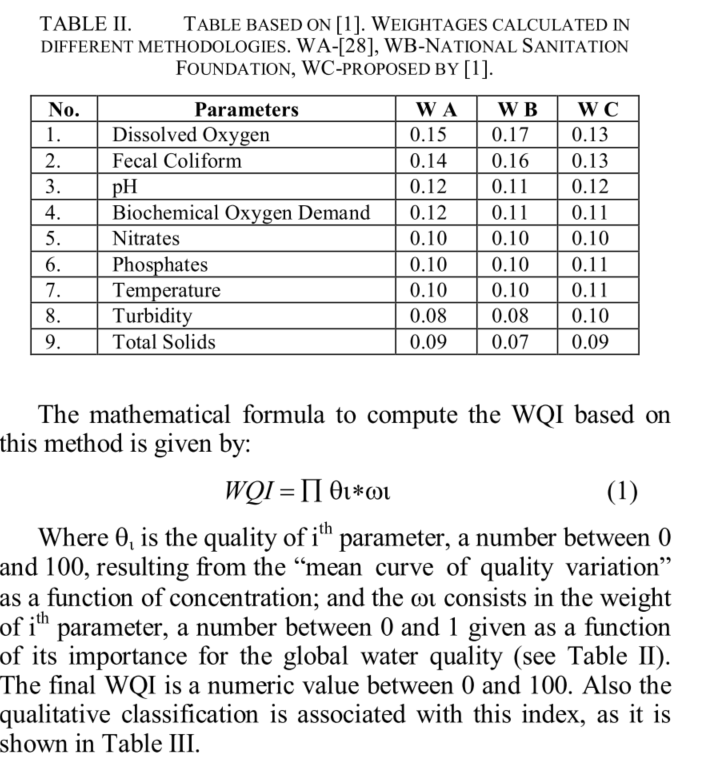
**Result:** kNN performed slightly better

**Gap: 1716 vectors from data, limited datasets, Prediction on only a narrow range**

**Parameters:** pH, Dissolved Oxygen, Nitrates and Temperature

ML Algo used:





**Data analysis, quality indexing and prediction of water quality for the management of rawal watershed in Pakistan (13 citations) (bad paper)**

**Ali & Qamar(2013)**

**Ali, M., & Qamar, A. M. (2013, September). Data analysis, quality indexing and prediction of water quality for the management of rawal watershed in Pakistan. In *Eighth International Conference on Digital Information Management (ICDIM 2013)* (pp. 108-113). IEEE.**

**(watershed)**

**Aim:** Forecasting water quality parameters based on trends.Water quality indices were analyzed using supervised and unsupervised ML techniques. The trend of the data was used to make predictions

**Technology/Algorithm:**  unsupervised learning technique of Average Linkage (Within Groups) method of Hierarchical Clustering using Euclidean distance

**Result:** both the supervised and unsupervised algorithm performed well and reasons for high contamination and source for fecal coliform was found

**Gap:** month wise data, limited datasets

**Parameters:** Appearance, Alkalinity, Temperature Hardness as CaCO3 , Conductance, Calcium, Total Dissolved Solids, Chlorides, Nitrite as NO2, Turbidity, pH and Fecal Coliforms

**An evaluation model of water quality based on DSA-ELM method**

**Yan et al. (2017) (2 citation)/ Bad paper**

**Yan, H., Liu, Y., Han, X., & Shi, Y. (2017, August). An evaluation model of water quality based on DSA-ELM method. In *2017 16th International Conference on Optical Communications and Networks (ICOCN)* (pp. 1-3). IEEE.**

**(River)**

**Aim:** Water Quality was classified using Extreme Learning Machine Algorithm optimized by Dolphin Swarm Algorithm (DSA-ELM)

**Technology/Algorithm:**  Ph, dissolved oxygen(DO), Potassium permanganate index (CODMn) and ammonia-nitrogen(NH3-N)

**Result:** Good accuracy and stability was seen (83.33%)

**Gap:** 150 groups of data, low sample size

**Parameters:**  Ph, dissolved oxygen(DO), Potassium permanganate index (CODMn) and ammonia-nitrogen(NH3-N)

**Surface water pollution detection using internet of things.  
Shafi et al. (2018) (5 citations) (Not great)**

**(different water sources)  
Shafi, U., Mumtaz, R., Anwar, H., Qamar, A. M., & Khurshid, H. (2018, October). Surface water pollution detection using internet of things. In *2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT)* (pp. 92-96). IEEE.**

**Aim:** Proposed IoT based solution to remote monitor water quality in real time (3 parameters) and control flow remotely using mobile app. It also classifies the water quality based on WHO guidelines on safe ranges and trained ML models with 9 parameters

**Technology/Algorithm:** Support Vector Machine (SVM), k Nearest Neighbor (kNN), single-layer neural network and deep neural network were used to classify the water quality and deep neural network

**Result:** Deep Neural Network outperformed all with accuracy of 93% for classification.

**Gap:** 667 samples from 11 different sources. Limited samples

**Parameters :** pH, Turbidity, Hardness as CaCo3, Conductance, Alkalinity, Dissolved Solids, Nitrate ,Fecal Coliform , Calcium for Machine Learning and

Only pH, Turbidity & Temperature Sensor for the IoT module

**Retrieval of Case 2 Water Quality Parameters with Machine Learning.**

**Ruescas et al. (2018) (2 citations)**

**Ruescas, A. B., Mateo-Garcia, G., Camps-Valls, G., & Hieronymi, M. (2018, July). Retrieval of Case 2 Water Quality Parameters with Machine Learning. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium* (pp. 124-127). IEEE.**

**(Northern Coastal Baltic Sea & Inland waters in the cold boreal zone)**

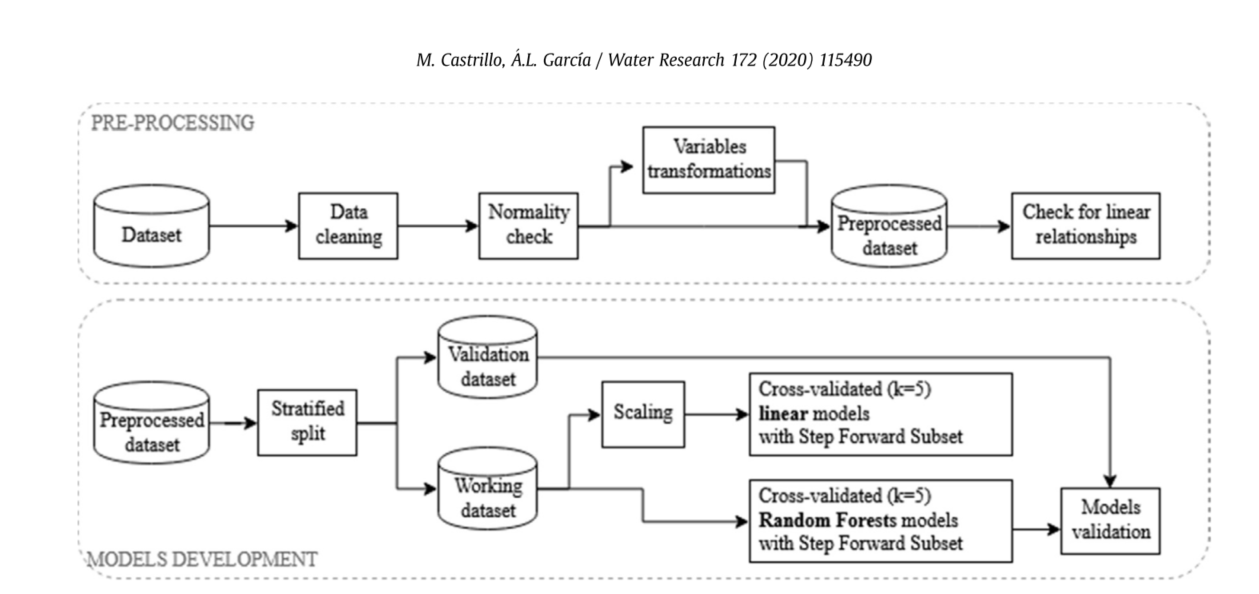
**Aim:** ML algorithms were used to retrieve Water quality parameters that were trained using simulated remote sensing reflectance. The ML algorithm maps reflectances to WQ parameters.

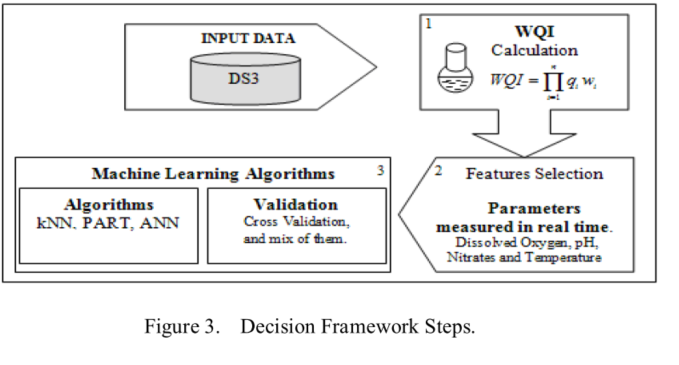
**Technology/Algorithm:** multivariate linear regression (RLR), random forest regression (RFR), Kernel ridge regression (KRR), Gaussian process regression (GPR) and support vector machine regression (SVR)

**Result:** Chl-a & CDOM, RFR and GPR performed best. RLR is not best for any case but is good enough for TSM

**Gap:** Reflectance data, expensive sensors

**Parameters:** Retrieved: Colored Dissolved Organic Matter (CDOM), Total Suspended Matter (TSM), Chlorophyll-A





**REFERENCES**

Ali, M., & Qamar, A. M. (2013, September). Data analysis, quality indexing and prediction of water quality for the management of rawal watershed in Pakistan. In *Eighth International Conference on Digital Information Management (ICDIM 2013)* (pp. 108-113). IEEE.

Camejo, J., Pacheco, O., & Guevara, M. (2013, January). Classifier for drinking water quality in real time. In *2013 International Conference on Computer Applications Technology (ICCAT)* (pp. 1-5). IEEE.

Cao, S., Wang, S., & Zhang, Y. (2018, December). Design of River Water Quality Assessment and Prediction Algorithm. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 901-906). IEEE.

Joslyn, K., & Lipor, J. (2018, December). A Supervised Learning Approach to Water Quality Parameter Prediction and Fault Detection. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 2511-2514). IEEE.

Khan, Y., & See, C. S. (2016, April). Predicting and analyzing water quality using Machine Learning: A comprehensive model. In *2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT)* (pp. 1-6). IEEE.

Muharemi, F., Logofătu, D., & Leon, F. (2019). Machine learning approaches for anomaly detection of water quality on a real-world data set. *Journal of Information and Telecommunication*, *3*(3), 294-307.

Ruescas, A. B., Mateo-Garcia, G., Camps-Valls, G., & Hieronymi, M. (2018, July). Retrieval of Case 2 Water Quality Parameters with Machine Learning. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium* (pp. 124-127). IEEE.

Shafi, U., Mumtaz, R., Anwar, H., Qamar, A. M., & Khurshid, H. (2018, October). Surface water pollution detection using internet of things. In *2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT)* (pp. 92-96). IEEE.

Wang, Y., Zhou, J., Chen, K., Wang, Y., & Liu, L. (2017, November). Water quality prediction method based on LSTM neural network. In *2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)* (pp. 1-5). IEEE.

Yan, H., Liu, Y., Han, X., & Shi, Y. (2017, August). An evaluation model of water quality based on DSA-ELM method. In *2017 16th International Conference on Optical Communications and Networks (ICOCN)* (pp. 1-3). IEEE.