A STUDY ON MACHINE LEARNING APPROACHES FOR WATER QUALITY DETECTION

[1] Muskan Bhalawat,[2] Anima Sharma

[1] B.Tech Research Scholar,[2] Assistant Professor,

[1,2] Department of Computer Science , Jaipur Engineering College and Research Centre ,Jaipur ,Rajasthan.

Abstract : Water is an inorganic , transparent , tasteless , odourless chemical substance , which is the one of the main constituents of the Earth's hydrosphere and is significant for all vital forms of life but Rapid urbanization has result in its deterioration . Predicting recreational water quality in inexpensive ways is one in every of the foremost difficult task and so as to try to do that the models and algorithms of machine learning were explored by the researchers . The model includes artificial (ANN) , deep (DNN) , recurrent neural network (RNN) , linear discriminant analysis (LDA) , super vector machines (SVM), logistic regression and long-short term memory (LSTM) while the algorithms checks the water quality index (WQI) and also the water quality class (WQC) . The Simulation study is conducted to test performance of every algorithm using F-score metric . The study was conducted so as to analyse the algorithms and methods together and can further provide the sunshine to the trail of future research on the water quality detection using advanced techniques.

***IndexTerms* – Machine Learning , Water Quality Detection , Deep Learning .**

I. INTRODUCTION

Water makes about 70% of the Planet i.e..2/3 of total surface and its vital for all kind of life sustaining here however life itself is becoming the threat to water . The surging , tormenting diseases are the results of poor water quality . The pollution results in Parasitic , Bacterial and Viral diseases such as typhoid , cholera , encephalitis , poliomyelitis , hepatitis , etc. which had became an acute problem especially in developing Countries .

The methods used before applying these machine learning algorithms and models were quite expensive and time consuming because those methods includes sample collection , transportation to labs and statistical analysis . Therefore the basic requirement was of the adequate and accurate agitating system that enables for simple and early recognization of all the changes , in order to take the required necessary decisions at the proper time . The major challenge was that of the noise and the highly imbalanced data because of which the prediction of anything was becoming a herculean task . Therefore the investigation was exhausted in order to look at the validity of popular algorithms on the highly imbalanced data and therefore the comparision was made between statistical and machine learning algorithms . The experiment was performed within the phases so as to grasp that which one could gain the simplest result . Within the first phase the classification of information was performed through the statistical algorithm logistic regression and therefore the second part was entended to the machine learning techniques like ANN , DNN,RNN , LDA , SVM , LSM . After these a representative set of supervised machine learning Algorithm were employed on dataset for predicting the water quality index and Sophistication . It had been concluded that Machine learning algorithms can simply deduce the quantity of false predictions and achieved good results for anomaly detections .

II. LITERATURE REVIEW

Byer and Carlson [1] were the first to form and test an internet monitoring of portable drinking water distribution system .

Zhang , Zhu , Yue and Wong [2] proposed a novel anomaly detection algorithm for water quality data using dual time moving windows.

Mohammadpour el al. [3] have investigated the matter by using super vector machines and a couple of methods of artificial neural network .

Kang , Gao and Xie [4] developed model for water quality prediction , the simplest result was achieved using artificial neural network with non linear autoregressive.

Recurrent Neural Networks and Long-Short term memory are pretty good at extracting patterns in input feature space where the input file spans overlong sequence. The estimated water quality in work is predicted on only three parameters : Turbidity , Temperature and pH , tested in step with the WHO.

Ahmat et al. [5] employed single feedforward neural network [NN] and a mix of multiple neural networks to estimate water quality index [WQI] .

Sakizadeh [6] predicted the water quality index [WQI] using 16 water quality parameters and Artificial Neural Network [ANN] with Bayerian regularization .

Rankonc et al. [7] predicted dissolved Oxygen using feedforward Neural network with 10 parameters .

Gazzaz et al. [8] used artificial neural network [ANN] to predict water quality issue [WQI] but the parameters included were 23 in number which made it quite expensive .

Ali Qamar [9] classified samples into quality classes but he ignored major parameters .

Abyaneh [10] predicted Chemical Oxygen Demand [COD] and Biochemical Oxygen Demand [BOD] using Artificial Neural Network [ANN] and multivariate rectilinear regression parameters utilised by him were pH , Temperature , Total Suspended , Total Suspended Solids .

Hamza Ahmad Isiyaka et al. [13] inspected and initiated a decrease in quantity of WQ monitoring places and gave the best input mixture for the WQ modelling utilising artificial neural network [ANN].

A.K. Bisht et al. [14] performed an task in order to access and model the WQ of several rivers utilising the method of Artificial Neural Network [ANN].

Salisu Yusaf et al. [15] classified the quality of water as well as prepared a suitable classification model for the analysis .

Muharemi et al .[16] proposed Neural network of classification based on logistic regression and Nearest Neighbour Algorithm [kNN].

III. METHODOLOGY

The real world data has noice and is extremely inconsistent , so to use the information for any work, we need to follow some steps , these can be described as

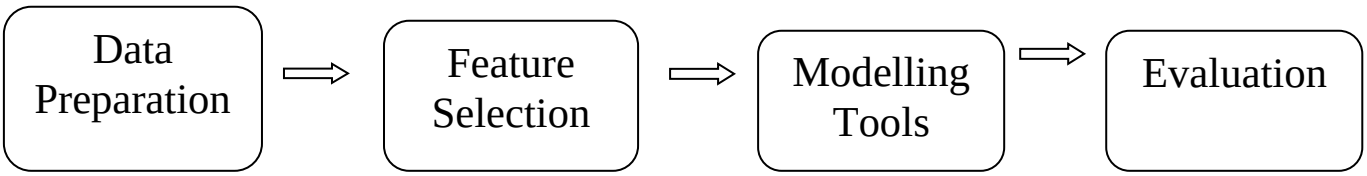

Fig 1 Process Flow

- *Data preparation* : It is must so as to induce the precise and accurate result . Some important steps within the data preparation are –
  - Data Cleaning : It is the method of detecting and correcting corrupt and inaccurate records from a record set and replacing or deleting the coarse data .
  - Data Transformation : It is the process of converting data from one format or structure into an another one .
  - Data Integration : It involves combining data residing in several sources and providing with a unified view of them.
  - Data Normalization : It means to scale a variable to own values between 0 and 1 .
  - Noise Identification : Noise may be a great deal of meaningless information ,which affects the data collection.

- *Feature Selection* : The amount of input variables when developing a predictive model are being reduced during this process which in successively reduces the computational cost of modelling and may even improve the performance of the model. The algorithm used for the feature selection were the
  - Embedded method : It complete the feature selection process within the development of the machine learning algorithm itself .
  - wrapper method : It works by the process of evaluating the subset of the features using the machine learning algorithm one that employs an exploration strategy in order to appear through the space of possible feature subsets , evaluating each subset supported the standard of the performance of a given algorithm .

  and the strategies used showed the similar result .

- *Modelling Tools :* , The method of modelling means to coach the machine learning algorithms to predict the labels from the feature , tuning it as per the requirement and validating it on holdout data.

- *Evaluation :* learning imbalanced data may be a challenging problem of data mining. To evaluate the performance Precision , Recall and F-score were taken into consideration.
  - Precision refers to the total of correctly classified positive examples divided by the amount of examples that were labelled as positive by the system , i.e..
    - P = TP / (TP+FP)
  - Recall is refered to as the total of correctly identified positive examples divided by the sum total of positive examples
    - Recall = TP / (TP+FN)
  - F-score is the harmonic mean of Precision and Recall .
    - F1 = 2* (Precision * Recall) / (Precision + Recall)

The major focus of them was to extend the F1 score.

The Another methodology adopted was :

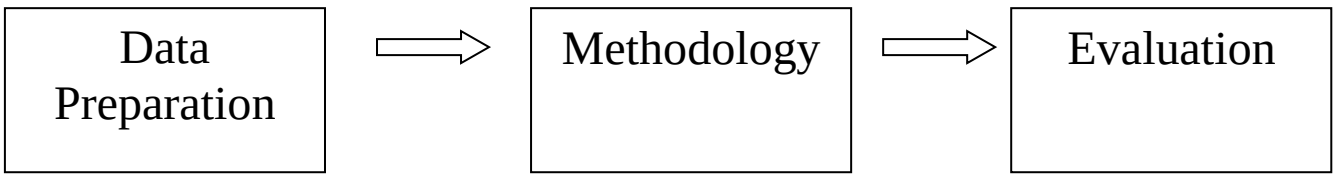

Fig 2. Other Methodology

- *Data Preparation :* included data collection , outlier Dtection , z-score normalisation .
  - Data Collection : It is about collecting the samples and listing down the parameters .
  - Outlier detection : It was done with the assistace of boxplot Analysis because it provides us with good visualisation with threshold detection .

- *Methodology :*

  - Water Quality Index[WQI] is a measure of water quality calculated using various parameters

    $WQI = (\sum Q_{value} * W\text{-factor}) / \sum W\text{-factor}$

    Where W-factor is the weighing factor .

    Parameters for water quality index calculation were :

Table 1. Parameters for Water Quality Index

Weighing Factor	Weight
Ph	0.11
Temperature	0.10
Turbidity	0.08
Total Dissolved Values	0.07
Nitrates	0.10
Fecal Coliform	0.16

  - Water Quality Class [WQC] is the range in which water quality should vary and is caluculate after knowing the WQI only

Table 2 Water Quality Class

Water Quality Index Range	Class
0 – 25	Very bad
25 – 50	Bad
50 – 70	Medium
70 – 90	Good
90 – 100	Excellent

- *Evaluation :*
  - Normalization of parameters was required , and so as to perform normalization , the methods used were –
    - Q-value Normalization was used to fit the parameters of water within the range of 0-100.
    - Z-score Normalization represents the amount of standard deviations and is conventional normalization and standardisation method.

      $Z\text{-score} = (x-\mu) / \sigma$

      Where , X = value of sample , μ = mean , σ = Standard deviation
- *Data Analysis :* By applying Machine learning algorithms and using minimal numbers of parameters data analysis was done for WQI and WQC predictions . The methods adopted for the Analysis were

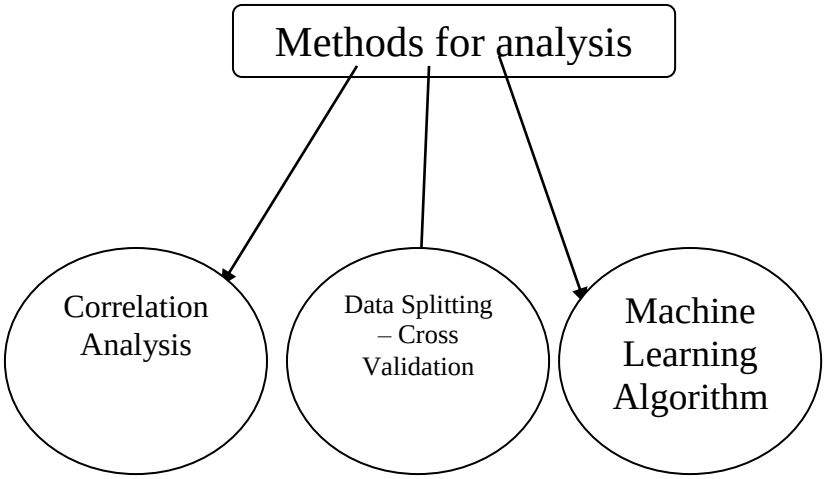

Fig 3 Methods for Analysis

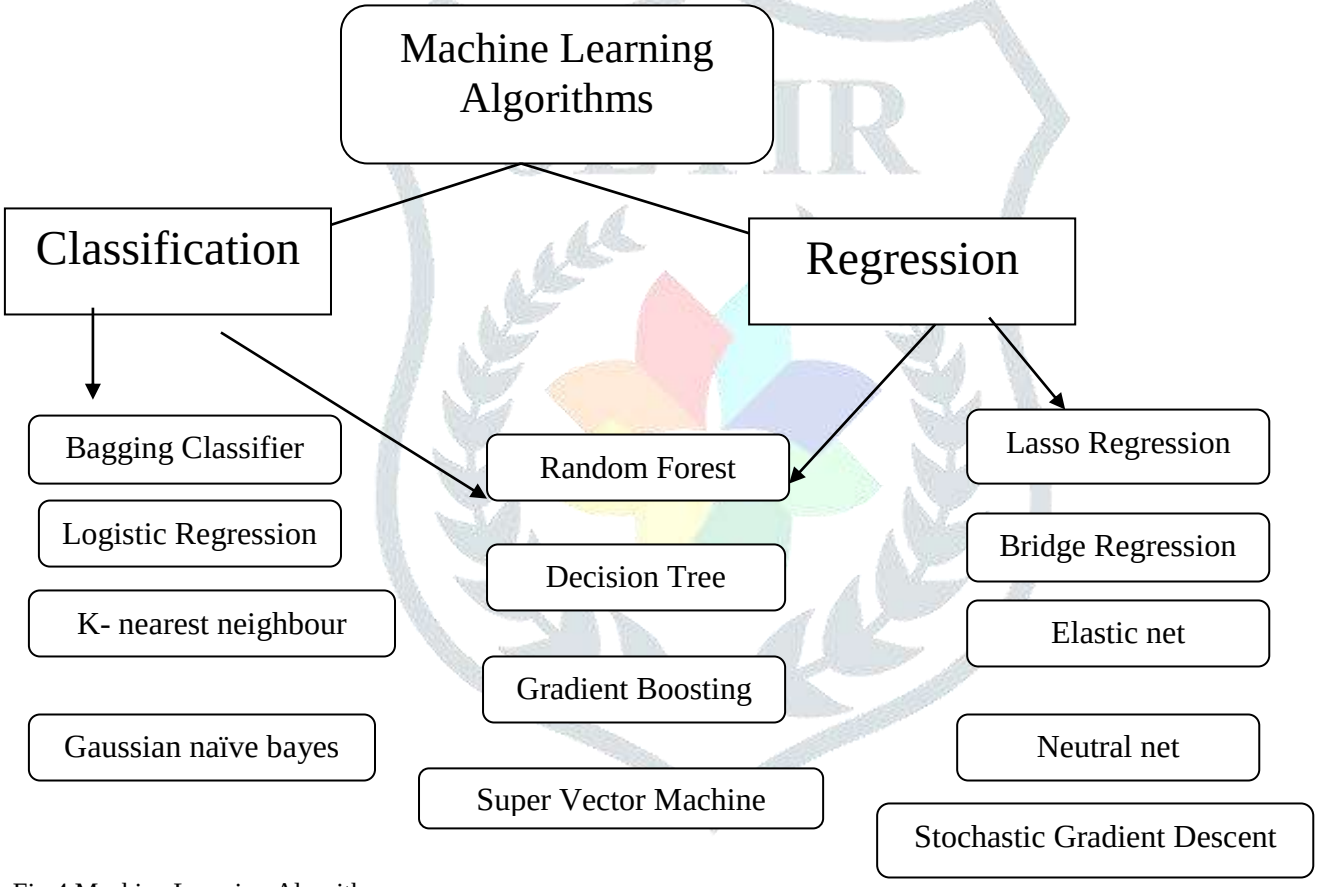

Fig 4 Machine Learning Algorithms

IV. Results

The result concluded by [11] was stated as the major focus area of the work done was to find the most efficient model for water quality data and all the algorithms used were unfortified some less (like SVM , ANN and Logistic Regression) and the others at the high level(like DNN , RNN , LSTM). The greater the F1 score the simpler was the result .

The F1 Score for the models as calculated was –

SVM	0.36
DNN	0.06
LSTM	0.146
RNN	0.1
LogRegression	**0.44**
Simple NN	0.32

The another study was done to calculate WQI and WQC [12] . In this the result was calculated checking each term for some accuracy measures.

Results for Regression algorithm calculated using

- Three parameters concluded Gradient Boosting as the simplest out of all the others with
  - MAE 1.9642 (least) ,
  - MSE 7.2011 (least) ,
  - RMSE 2.6835 (least) ,
  - R squared 0.7485 (max).
- Four parameters concluded Polynomial Regression as the simplest out of all with
  - MAE 2.7273 (least) ,
  - MSE 12.7307 (least) ,

So out of the all Logistic Regression gave the best result	RMSE 3.5680 (least) , R squared 0.4851. So, Gradient boosting and polynomial regression was most efficient out of all for calculating WQI. Results for Classification algorithm calculated using four parameters concluded MLP as the simplest and efficient out of all with Accuracy 0.8507(max) , Precision 0.5659(max), Recall 0.5640(max), F1 Score 0.5649(max).

V. Conclusion

In [11] the approaches for event detection on water quality time series data was presented . This was a case study conducted to find the most efficient model out of all the machine learning and statistical models for water quality detection .

But the imbalance in the data is of very high level , so it was concluded to find some more universal approaches for the same .

In [12] its main aim was to minimise the cost of all the expenses in the name of water quality detection at labs , a set of machine learning algorithms were employed for the detection of WQI . This was found useful , it deduced the cost of quality calculation and impoved the accuracy .

They concluded to apply it for the real time set of data using IOT which will have different sensors representing water parameter which will let the people drink pure water and will subsequently reduce the chances of water born disease

References

1. Byer , D.,& Carlson , K.H. in the year 2005 in Journal-American water works association 97(7) named “Real time detection of international chemical combination in the distribution system “.
2. Zhang , J.,Zhu , X., Yue , Y., & wong P.W. in the year 2017 in seventh international conference on innovative computing international technologies (INTECH) (pp . 36-41). IEEE named “ A real time anomaly detection algorithm/ or water quality data using dual time moving windows”.
3. Mohammadpour , R.,Shaharuddin,S.,Chang , C.K.,Zakaria,N.K.,Ab Ghani , A.,and Chan,A.W. in the year 2015 in Environmental Science and pollution research , 22 (8) , 6208-6219 named “Prediction of water quality index in constructed wetlands using support vector machine”.
4. Kang ,G.K., gao,J.Z. & xie G.(n.d.) conducted a survey named “Data driven water quality analysis and prediction”.
5. Ahmad,A ; Khalid,N. ; Javed , T. in 2013 named” Drinking water quality in capital city of Pakistan” .
6. Sakizadeh in 2016 in Model Earth Syst Environ 2,8 named “Artificial Intelligence for the prediction of WQI in ground water system” .
7. Rankovic , v.; Radulovic , J; Radojevic,I.; Ostojic , A.; Comic,L. , Serbia ecol model 2010 , 221 , 1239 – 1244 named “ Neural Network modelling of dissolved oxygen in the gruza reservoir”.
8. Gazzaz , N.M.;Yussof , M.K.; Aris , A.Z .; Juahir H.; Ramli , M.F. in 2012 in Mar. pollot. Bull , 64 , 2409-2420 named “Artificial neural network modelling of the water quality index for kinta river (Malaysia) using water quality variables as preditors” .
9. Ali , M.; Qamar , A.M. on 10-12 september , 2013 , pp. 108-113 in proceedings of English International Confrence named “ Digital Information Management on Data Analysis , quality indexing and prediction of water quality for the management of rawal watershed in Pakistan” .
10. Abyaneh , H.Z. in 2014 in J. environ. Health Sci. Eng. 12,40 named “Evaluation of multivariate linear regression and artificial neural network in prediction of water quality parameters “.
11. Fitore Muharemi , Doina Logofatu and florin leon in 2019 in journal of Information and telecommunications named “Machine learning approaches for an aomaly detection of water quality on real world dataset”.
12. Umair Ahmed , Rafia Mumtaz , Hirra Anwar , Asad A. Shah , Rabia Irfan & Jose Garcia-Nieto , in 2019 named “Efficient water quality prediction using supervised Machine learning” .
13. Hamza Ahmad Isiyaka , Adamu Mustapha , Hafizan Juahir and Philip Ogbonnia Phil – Eze, in 2018 in Modelling Earth System and Environment pp. 1-11 named “Water Quality Modelling using Artificial Neural Network and multivariate Statistical techniques”.
14. A.K. Bisht , Dr. Ravendra Singh , Rakesh Bhutiani and Ashutosh Kumar Bhatt in 2017 on International Journal on Future Revolution in Computer Science and Communication Engineering named “ An Investigation of Assessment and Modelling the Water Quality of Rivers Based on Artificial neural network - An initiative towards river Ganga”
15. S.Y.Muhammad , M.Makhtar , A. Rozaimee , A. A. Aziz , A. A. Jamal in 2015 in International journal of Software engineering and its applications , pp 45-52 named “ Classification Model for water Quality Using Machine Learning Techniques “ .
16. Muharemi , F. , Logofatu , D., Andresson ,C., & Leon , F. named “ A Case Study on Approaches to build a decision model for water quality.
17. Che ,Z.,Purushotam,S.,Cho,K.,Sontag,D., & Liu , Y. in 2018 in Scientific Reports (2018) named “Recurrent Neural Network for Multivariate Time Series with Missing Values ”.

18. Sapankevych,N.I.,& Sankar R., in 2009 in IEEE Computational Intelligence Magazine , 4(2) 24-38 named "Time Series Prediction using Support Vector Machine ".
19. Sokolova, M., & Lapalme, G., in 2009 in Information Processing & Management , 45(4) , 427-437 named "A Systematic Analysis of Performance Measure for Classification Task".
20. Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. in 2008 in preceedings of 25th International Confrence on Machine Learning (pp . 1096-1103) ACM . "Extracting and Composing Robust Features with Denoising Autoencoders ".