

USING MACHINE LEARNING AND IMAGE RECOGNITION TO EVALUATE WATER QUALITY

Anil Bellapu ¹, * Dr. Dileep kumar Padidem ², T Lakshmi Prasanna ³

BI Engineer/Solution Architect, Costco Wholesale

² Professor, Ramireddy Subbarami Reddy Engineering College, Kavali, A.P.- 524201_

³ Assistant Professor, Narayana Engineering College, Nellore, A.P.-524004

*Corresponding Author

Email: padidemdileep@gmail.com

Abstract – This research project explores the application of machine learning in assessing water quality through image recognition. The study leverages a diverse dataset of water samples collected from various sources across Mumbai, encompassing ponds, lakes, water outlets, and sewage points. Multiple parameters, including pH, conductivity, turbidity, and dissolved oxygen content, are examined in relation to the resulting HEX colour code, which serves as a visual representation of water quality. The central hypothesis posits that employing machine learning algorithms can reliably predict water safety based on these environmental parameters through image recognition of the colour gradient of the water samples. A systematic approach to data collection, standardisation, and logistic regression modeling has been employed. Results demonstrate the effectiveness of the logistic regression model in predicting water safety with an 85% accuracy rate, highlighting its potential for real-time water quality monitoring and risk assessment. Nevertheless, this study recognises its limitations and the need for further research to refine the model's predictive accuracy and address variations across different geographical regions and water sources. This research aims to contribute to the development of innovative approaches for water quality assessment and therefore, environmental preservation.

Key Words: Machine learning, water sources, pH, conductivity, turbidity, dissolved oxygen, HEX colour code, logistic regression

1. INTRODUCTION

Clean water and adequate sanitation are vital for human existence, environmental sustainability, and economic development. However, the irresponsible disposal of industrial effluence has emerged as a significant threat to water quality in major world cities including Mumbai, a particularly urbanized region. This research paper aims to investigate how pollutants—measured through various water quality parameters such as pH, turbidity, dissolved oxygen and conductivity—affect the colour gradient of various water bodies and therefore, the overall water quality.

1.1 PROBLEM STATEMENT

While water quality using the aforementioned parameters (pH, conductivity, turbidity, and dissolved oxygen) has been tested and analysed in the past, this research paper looks into the feasibility of using image recognition and machine learning as a way of predicting water quality - in essence using technology to look at a very widely prevalent, yet fundamental community problem. The fusion of machine learning with water quality assessment has immense potential to make water safety evaluations more widely accessible.

2. BACKGROUND INFORMATION

The World Wildlife Fund (WWF) defines water pollution as “toxic substances entering water bodies such as lakes, rivers, oceans and so on, getting dissolved in them” (“Water Pollution”).

EFFECT OF WATER ON FOUR WATER QUALITY PARAMETERS

Effect of Water Pollution on Electrical Conductivity: Electrical conductivity is a measure of how well a substance allows the flow of electric charges, typically in the form of electrons or ions.

Water pollution often introduces various charges into aquatic ecosystems. For example, agricultural runoff, industrial discharges, and urban runoff can contain ions like sodium, chloride, nitrate, sulfate, and heavy metal ions. These ions increase the electrical conductivity of water as they are capable of carrying electrical charges [5].

Importantly, there is a relationship between pH and electrical conductivity. Heavy metals found in water pollution affect pH (discussed in earlier paragraphs), and consequently, these changes in pH can affect the dissociation of ions, potentially impacting electrical conductivity. Acidic conditions, for example, can lead to the release of hydrogen ions (H^+), which can alter the conductivity.

Effect of Water Pollution on Turbidity:

Turbidity is a measure of the cloudiness or haziness of a fluid caused by the presence of suspended particles, which scatter and absorb light. Turbidity is an essential parameter in environmental and water quality monitoring, as it can indicate the presence of contaminants or impurities in a liquid [3].

Pollutants such as sediment, clay, organic matter, and debris can be introduced into water bodies through various sources, including erosion, industrial discharges, and sewage runoff. These particles can increase turbidity, making the water appear cloudy or visibly spoilt. Industries that release effluents containing fine particles, chemicals, heavy metals, and other contaminants can significantly increase turbidity in nearby water bodies. Construction sites in Mumbai, are also the cause of such particles entering water bodies (Tembhekar et al.).

A point to note is that with greater suspended particles in a water sample, there is a higher chance of finding metals in the water. This is notable because with a higher amount of suspended particles, the impurities containing ions allow electricity to be easily conducted. Hence, turbidity and electrical conductivity have a direct relationship, wherein, the increase in the value of one will result in the increase of the other.

Effect of Water Pollution on Dissolved Oxygen:

INSTRUMENTS MEASURING PARAMETERS

1. pH Meter - pH: scale from 1-14
2. Conductivity Meter - Electrical Conductivity: expressed in microsiemens per centimeter ($\mu S/cm$).
3. Turbidimeter - Turbidity: expressed in Nephelometric Turbidity Units (NTU) or Parts Per Million (PPM).
4. Nephelometer (DO meter) - Dissolved Oxygen: expressed in milligrams per liter (mg/L) or parts per million (ppm)

HOW PARAMETER AFFECTS THE COLOUR

pH can influence the colour of water indirectly by affecting the solubility and speciation of certain substances. For example, water with a low pH (acidic) can cause the leaching of metals like iron and manganese, leading to discolouration in the form of reddish-brown or yellowish tints. Extreme pH values can also impact the colour perception, making the water appear more or less translucent.

Conductivity measures the concentration of ions in water, mainly due to the presence of salts and minerals. While high conductivity can be indicative of dissolved ions, it does not influence the colour of the water directly. High ion concentrations can affect the taste and odor of water, but they do not define the colour.

Turbidity is a measure of the cloudiness or haziness of water due to suspended particles. The presence of suspended particles in water can scatter and absorb light, leading to variations in the water's appearance.

High turbidity can make water look cloudy or murky, which can affect the perceived colour.

The dissolved oxygen content does not have a direct impact on the colour of water. However, it can indirectly affect the health of aquatic ecosystems and the growth of algae or other organisms that may impart different colours to water, such as green or brown hues. Changes in dissolved oxygen levels can influence the biological activity and the presence of organic matter in water, which may impact the water's overall appearance.

3. SCIENTIFIC LITERATURE REVIEW

Groundwater quality is assessed through various parameters such as turbidity, pH, dissolved oxygen (DO), and conductivity. Recent studies have also explored the application of machine learning in predicting and managing water quality. This literature

review categorizes the selected research into these themes to provide a cohesive understanding of the topic.

4. VARIABLES

Independent Variables:

The independent variables in this study are represented by the various water bodies from which water samples are procured. These water bodies serve as distinct categories or groups for the machine learning algorithm's analysis. Within each category, the parameters, including pH, conductivity, turbidity, and dissolved oxygen content, are measured and act as the features or characteristics used for the machine learning algorithm. These parameters are the focus of the study particularly concerned with their relationship to the HEX colour code, and the different water bodies, from different locations, are the independent variable..

5. HYPOTHESIS

The utilization of a machine learning algorithm for predicting water safety based on the parameters of pH, conductivity, turbidity, and dissolved oxygen content, within various water bodies, will result in a reliable and efficient method for assessing water quality, as represented by the corresponding HEX colour code.

Supporting Evidence:

pH levels influence the colour of water indirectly by affecting the solubility and speciation of certain substances. For instance, low pH (acidic) can lead to the leaching of metals

like iron and manganese, which can result in reddish- brown or yellowish tints, thus affecting the HEX colour code. Conductivity measures ion concentration in water, impacting its taste and odor. High conductivity does not directly affect the HEX colour code, but it can indirectly influence water quality which in turn may have an effect on the colour composition of water. High ion concentrations can alter water characteristics, potentially impacting the HEX colour code. **METHODOLOGY**

The research employed a systematic approach to collect and analyse water samples from diverse sources across Mumbai and develop a logistic regression model for water quality assessment.

1. Sample Collection:

60 water samples were collected from various locations, including ponds, lakes, water outlets, and sewage sources throughout Mumbai, to ensure a broad spectrum of water sources and conditions. The collection was carried out using a bucket and a rope. Each sample was collected in individual sample cups to maintain sample integrity. 3-5 samples of the same source were collected.



Fig.6.1-6.6: Pictures of Various Ponds/Industrial Sites Across Mumbai from which Samples were Collected

2. Data Collection:

Multiple water quality parameters were measured for each collected sample. The IONIX 7 in 1 multiparameter instrument was used to gather data on pH, electrical conductivity, and turbidity. The Lutron DO-5509 Dissolved Oxygen Meter was used to collect data for dissolved oxygen. Both instruments were calibrated and standardized before use. Three trial measurements for each sample were taken.



Fig.6.7:DataCollectionusingaDigitalMultiparameter and a Nephelometer

3. Data Analysis & Standardization: The collected data, including pH, conductivity, turbidity, and DO values, was transferred to an Excel spreadsheet for analysis. This data was organized and averaged to prepare it for the development of a logistic regression model. Standard values for drinking water from secondary sources, mentioned in the background research, allowed for a

Sample Collection:

60 water samples were collected from various locations, including ponds, lakes, water outlets, and sewage sources throughout Mumbai, to ensure a broad spectrum of water sources and conditions. The collection was carried out using a bucket and a rope. Each sample was collected in individual sample cups to maintain sample integrity. 3-5 samples of the same source were collected.



Fig.6.1-6.6: Pictures of Various Ponds/Industrial Sites Across Mumbai from which Samples were Collected

4. Data Collection:

Multiple water quality parameters were measured for each collected sample. The IONIX 7 in 1 multiparameter instrument was used to gather data on pH, electrical conductivity, and turbidity. The Lutron Do-5509 Dissolved Oxygen Meter was used to collect data for dissolved oxygen. Both instruments were calibrated and standardized before use. Three trial measurements for each sample were taken.



Fig.6.7:DataCollectionusingaDigitalMultiparameter and a Nephelometer

Data Analysis & Standardization: The collected data, including pH, conductivity, turbidity, and DO values, was transferred to an Excel spreadsheet for analysis. This data was organized and averaged to prepare it for the development of a logistic regression model. Standard values for drinking water from secondary sources, mentioned in the background research, allowed for a human decision on whether the water was potable or contaminated beyond saving. This was indicated by a binary factor: if the water sample's data was within the standard drinking water ranges, it would yield a potability factor 1; however, if the data reflected digits outside of the standard drinking water range, a potability factor 0 was attained. The two factors that were given highest priority were conductivity and turbidity since they showed a positive correlation, hence, the portability factor was based off of the standard drinking water values for these parameters

Sample No.	Average pH	Average turbidity	Average conductivity	Average DO	Hex code w/o#	Potability (0=nonpotable, 1=potable)
1	6.95	5327	10623	4.7	9a9e9a	0
2	6.99	5329	10610	4.7	8e9d8b	0
3	6.99	5317	10610	4.5	6b758e	0
4	6.99	5317	10617	3.9	81827e	0
5	6.99	5317	10603	3.9	9d918c	0
6	6.95	3267	6593	3.45	9c9e97	0
7	6.94	3283	6593	4.1	979690	0
8	6.93	3298	6593	3.5	747168	0
9	6.94	3298	6607	3.9	91928c	0
10	6.96	3299	6603	3.75	7d7c74	0
11	6.88	330	659	3.15	777b72	1
12	6.84	320	639	2.9	9d959b	1
13	6.85	320	639	2.6	777772	1
14	6.86	320	638	2.75	7b7971	1
15	6.85	320	639	2.25	5e5550	1
16	6.94	235	481	1.75	83847e	1
17	6.93	232	472	2.05	8b8c86	1
18	6.95	231	498	1.5	707071	1
19	6.95	252	501	2.35	8f8494	1
20	6.95	261	503	1.75	929797	1

Fig.6.8:Digitisationofdataandstandardizationof samples

Simultaneously, using the colour extraction algorithm the HEX code values and the respective intensity percentageof those HEX code values were extracted. These values were accounted for in the data sheet as they play a crucial role in the development of a supervised machine learning algorithm. Below is an example of one of the samples that underwent this process.



Fig.6.9:Colourexttractionofwatersamples

1. ModelDevelopment:

A logistic regression model was created using the processed data. Logistic regression is a type of regression analysis used to predict a dependent variable's outcome based on one or more independent variables. The dependent variable in logistic regression is binary, meaning it has only two possible outcomes (e.g., safe or unsafe water). Unlike linear regression, which predicts a continuous outcome, logistic regression predicts the probability of a certain class or event existing. In logistic regression, each feature (water quality parameter) is assigned acoefficient, which representsitscontribution to the logit (log-odds) of the outcome. The logit function is given by:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

Where p is the probability of the outcome being 1 (potable), β_0 is the intercept, β_i are the coefficients for the variables. The model coefficients (

where X_i are the independent variables) were estimated using the maximum likelihood estimation method. This involves finding the values of the coefficients that maximize the likelihood of an accurate model as per standard drinking water conventions. In other words, reducing the number of outliers.

	P(X=0)	P(X=1)	logit	e ^{logit}	Probability	log likelihood
b0	0.0000101031	-0.00001010398975	1.648074742	5.196964059	0.177125444	-0.9738320674
b1	0.0000106991	-0.00001069921962	1.327301918	3.770898853	0.200112062	-0.8417199194
b2	-0.000123692	0.0001236923435	0.09517462377	1.000997199	0.921510675	-0.08174091261
b3	-0.000247783	0.0002477836893	1.105977134	3.022176099	0.100100175	-0.7447795563
b4	0.0000100901	-0.00001009063405	1.598372371	4.94497770	0.166666667	-0.9548052377
b5	0.0000053681	-0.000005368638846	3.008841613	20.26391206	0.954989479	-1.388692325
b6	0.0000210611	-0.0000210613041	3.223187552	25.10892591	0.200100175	-1.440990172
b7	0.00001	-0.00001	1.972915174	7.19161076	0.200100175	-1.099643012
			2.892008084	16.04390734	0.954989479	-1.39913277
			1.970775409	7.176260133	0.200100175	-1.088024373
			3.973364031	53.16307256	0.100100175	-0.2244833797
			4.902162016	134.5831221	0.800100175	-0.1856417204
			3.970679101	53.46700063	0.200100175	-0.2242046389
			4.034311876	56.46741516	0.200100175	-0.2241814653

Fig.9-10: Set values for parameter that yield probabilities

The predicted probability of water safety (p) is then obtained by applying the logistic (sigmoid) function to the logit:

$$p = \frac{1}{1 + e^{-\text{logit}(p)}}$$

This converts the logit value into a probability ranging between 0 and 1. A threshold of 0.5 was used to classify the water as safe or unsafe. Probabilities greater than 0.5 were classified as potable (1), while probabilities less than 0.5 were classified as non-potable (0).

2. Model Development:

A logistic regression model was created using the processed data. Logistic regression is a type of regression analysis used to predict a dependent variable's outcome based on one or more independent variables. The dependent variable in logistic regression is binary, meaning it has only two possible outcomes (e.g., safe or unsafe water). Unlike linear regression, which predicts a continuous outcome, logistic regression predicts the probability of a certain class or event existing. In logistic regression, each feature (water quality parameter) is assigned a coefficient, which represents its contribution to the logit (log-odds) of the outcome.

The predicted probability of water safety (p) is then obtained by applying the logistic (sigmoid) function to the logit:

$$p = \frac{1}{1 + e^{-\text{logit}(p)}}$$

This converts the logit value into a probability ranging between 0 and 1. A threshold of 0.5 was used to classify the water as safe or unsafe. Probabilities greater than 0.5 were classified as potable (1), while probabilities less than

0.5 were classified as non-potable (0). **Fig.11: Sigmoid function (logistic regression curve)**

3. Model Validation:

Samples that show discrepancies between their probability (0 or 1) and their respective potable factor are highlighted in yellow and are considered anomalies in this model. The developed logistic regression model was tested and validated using an unknown sample and predicting its quality. An 85% accuracy rate was attained with the numericals of each parameter described above. These anomalies are highlighted in yellow as shown in Fig. 6.10.

6. RESULTS

The results of this study demonstrated the feasibility of employing a logistic regression model to assess water quality and predict potability based on water parameters within various sources. Water samples from a diverse range of locations across Mumbai, including ponds, lakes, water outlets, and sewage sources, were collected and analysed. Key water quality parameters, namely pH, conductivity, turbidity, and dissolved oxygen (DO), were measured using specialized instruments. Additionally, the HEX colour code values and their respective intensity percentages were extracted using a colour extraction algorithm, providing crucial information for the machine learning model.

4. Model Validation:

Samples that show discrepancies between their probability (0 or 1) and their respective potable factor are highlighted in yellow and are considered anomalies in this model. The developed logistic regression model was tested and validated using an unknown sample and predicting its quality. An 85% accuracy rate was attained with the numericals of each parameter described above. These anomalies are highlighted in yellow as shown in Fig. 6.10.

7. CONCLUSION

The findings of this research highlight the potential of utilizing a supervised machine learning algorithm, specifically a logistic regression model, to assess water quality based on key water parameters and the corresponding HEX colour code. The study leveraged water samples from diverse sources to ensure a comprehensive evaluation of water quality. pH, conductivity, turbidity, and DO were identified as critical factors for water quality assessment, as they influence the HEX colour code. The observed relationships between these parameters and the HEX colour code align with previous research findings, emphasizing the importance of these variables in determining water safety.

It is crucial to note that this research design and model rely on the relationships observed within the specific water bodies in Mumbai. Variations in geological and environmental conditions in other regions may influence the parameters' relationships differently. Hence, while this logistic regression model is promising for Mumbai, its generalizability to other locations necessitates further investigation.

Moreover, the findings of this study have practical implications for water quality management and public health. The ability to rapidly and accurately assess water safety using machine learning algorithms can facilitate timely actions to protect communities from potentially contaminated water sources. The implementation of this model in real-time monitoring and decision support systems could significantly enhance the safety of water supplies and improve the overall well-being of communities. Future research should focus on validating the model in diverse geographical and environmental contexts to ascertain its broader applicability.

While water quality using the aforementioned parameters (pH, conductivity, turbidity, and dissolved oxygen) has been tested and analysed in the past, this research paper looks into the feasibility of using image recognition and

machine learning as a way of predicting water quality - in essence using technology to look at a very widely prevalent, yet fundamental community problem. The fusion of machine learning with water quality assessment has immense potential to make water safety evaluations more widely accessible.

REFERENCES

- [1] US EPA, O. (2015, November 4). *Dissolved Oxygen*. *Www.epa.gov*. <https://www.epa.gov/caddis/dissolvedoxygen#:~:text=c%20The%20colour%20of%20water>
- [2] Electrical Conductivity and Electrical Resistivity - Resistivity of Materials, Formula, Unit, Examples and FAQs. (n.d.). BYJUS. Retrieved July 31, 2024, from <https://byjus.com/physics/resistivity-variousmaterials/#:~:text=Electrical%20Conductivity%20is%20an%20intrinsic%20property%20of%20a>
- [3] Huey, G. M., & Meyer, M. L. (2010). Turbidity as an Indicator of Water Quality in Diverse Watersheds of the Upper Pecos River Basin. *Water*, 2(2), 273–284. <https://doi.org/10.3390/w2020273>
- [4] Kang, C., Lee, T. M., Wong, B., & Yoo, J. (2018). Relationship Between Soil pH and Dissolved Oxygen (DO) Concentration in Salish Creek and Canyon Creek. *The Expedition*, 8. <https://ojs.library.ubc.ca/index.php/expedition/article/view/191429>
- [5] pH Measurement | Electrical Instrumentation Signals | Electronics Textbook. (n.d.). *Www.allaboutcircuits.com*. <https://www.allaboutcircuits.com/textbook/directcurrent/chpt-9/ph-measurement/>
- [6] Mathur, A. (2018). Conductivity: Water Quality Assessment. *International Journal of Engineering Research & Technology*, 3(3). <https://doi.org/10.17577/IJERTCONV3IS03028>
- [7] Pathak, A. K., Sharma, M., Katiyar, S. K., Katiyar, S., & Nagar, P. K. (2020). Logistic regression analysis of environmental and other variables and incidences of tuberculosis in respiratory patients. *Scientific Reports*, 10(1). <https://doi.org/10.1038/s41598-020-79023-5>
- [8] Water Science School. (2019, October 22). pH and Water | U.S. Geological Survey.