# MONITORING WATER QUALITY USING MACHINE LEARNING ALGORITHMS

[1] M.Vineela, [2] Basani Meghana, [3] Sathavarapu Harshitha
[1] Associate Professor, Department of CSE, BhojReddy Engineering College for Women, Hyderabad, Telangana, India.
[1] Vineela_m_99@yahoo.com
[2,3]Students, Department of CSE , BhojReddy Engineering College for Women, Hyderabad, Telangana, India.
[2] basani.meghanareddy@gmail.com , [3] harshithasathavarapu@gmail.com

*Abstract*—Monitoring water quality is a critical aspect of environmental sustainability. Poor water quality has an impact not just on aquatic life but also on the ecosystem. The purpose of this systematic review is to identify peer-reviewed literature on the effectiveness of applying machine learning (ML) methodologies to estimate water quality parameters with satellite data. The data was gathered using the Scopus, Web of Science, and IEEE citation databases. Related articles were extracted, selected, and evaluated using advanced keyword search and the PRISMA approach. The bibliographic information from publications written in journals during the previous two decades were collected. Publications that applied ML to water quality parameter retrieval with a focus on the application of satellite data were identified for further systematic review. A search query of 1796 papers identified 113 eligible studies. Popular ML models application were artificial neural network (ANN), random forest (RF), support vector machines (SVM), regression, cubist, genetic programming (GP) and decision tree (DT). Common water quality parameters extracted were chlorophyll-a (Chl-a), temperature, salinity, colored dissolved organic matter (CDOM), suspended solids and turbidity. According to the systematic analysis, ML can be successfully extended to water quality monitoring, allowing researchers to forecast and learn from natural processes in the environment, as well as assess human impacts on an ecosystem. These efforts will also help with restoration programs to ensure that environmental policy guidelines are followed

## I. INTRODUCTION

1.1. Water quality Water quality describes a state of a water body, as well as its chemical, physical, and biological aspects, including its usefulness for a particular activity (i.e., fishing, swimming or drinking). Substances that can damage aquatic species if found in high enough quantities can also impair water quality. Monitoring water quality is a critical aspect of environmental sustainability. Poor water quality has an impact not just on aquatic life but also on the ecosystem. The following variables are also be used to provide an indicator of water quality: the content of dissolved oxygen (DO); amounts of fecal coliform bacteria from people and animal wastes; levels or ratio of plant nutrients nitrogen and phosphorus; volume of particulate suspended matter (turbidity) and the amount of salt (salinity) in the water. To assess water quality, quantities of substances such as pesticides, herbicides, heavy metals, and other pollutants can be calculated. The abundance of chlorophyll-a (Chl-a), a green pigment present in microscopic algae, is often filtered from water samples in many water bodies to provide an indicator of the microalgae living in the water column [1].

1.2. Satellite and remote sensing Remote sensing is the method of surveying the surface of the earth without making any physical connection. It is used primarily to collect data from the earth's properties and analyze changes in the earth's environment. Along with improvements in satellite technologies and device processing capability, remote sensing has become more widely used in this era. Remote sensing generates spectral, infrared, and radar images that can be interpreted and analyzed to extract useful knowledge about earth elements like water, soil, plants, and the atmosphere, among others. These data are often used to forecast weather and environment, as well as for tracking animal populations, crop health, shoreline changes, and land-use change detection. The resolution of remote sensing data varies depending on the satellite capability. Remote sensing data has recently been produced and effectively utilized to collect water quality information as a solution to the limitations of traditional methods [2]. Remotely sensed data sets are usually more extensive than those collected directly on site by providing better resolution and typically higher temporal frequency and resolution for spatial coverage [3]. Remote satellite sensing examples include Landsat, Sentinel, MODIS, MERIS and VIIRS. 1.3. Machine learning Machine Learning (ML) is a type of statistical approach that can automatically learn from data and construct a detection, estimation, or classification model that minimizes the variance between the training and prediction datasets without being actively programmed. ML, also known as statistical learning, is providing data to a computer that can be "trained" using known or predetermined attributes or objects to allow semi-automatic or automatic detection, classification, or pattern recognition. ML enabling remotely sensed water quality estimate has grown in popularity in recent years as a result of improvements in algorithm development, computer power, sensor systems, and availability of data [4]. 1.4. Systematic review objectives In this systematic review, the effectiveness of applying ML methodologies were investigated to retrieve water quality parameters from satellite data. Specifically, the objective of studies, the types of satellite data, the ML methodologies, the significance or outcome of the ML application were summarized. 1.5. Nomenclature Figure 1 provided the list of the abbreviations, acronyms and symbols used in this manuscript.

## II. PROBLEM STATEMENT

According to the systematic analysis, ML can be successfully extended to water quality monitoring, allowing researchers to forecast and learn from natural processes in the environment, as well as assess human impacts on an ecosystem. These efforts will also help with restoration programs to ensure that environmental policy guidelines are followed.

## I. IMPLEMENTATION

present in microscopic algae, is often filtered from water samples in many water bodies to provide an indicator of the microalgae living in the water column [1]. The following sections provide a detailed overview of each of these steps.

**Step 1: Data Collection and Pre-processing:**

The first step in this research design is to collect and preprocess a dataset . The data is pre-processed by cleaning and normalizing the data and removing any duplicate or irrelevant information. Feature selection is also performed to select the most relevant features for the machine learning models. The eligibility of publications was evaluated and the publications were screened by examining the titles, abstracts and methods, and then obtained eligible publications through reading the full text.

**Step 2: Model Selection and Implementation:** The next step in this research design is to select and implement several machine-learning models. The models selected include Support Vector Machines (SVM), Naive Bayes, Decision Trees, Logistic Regression, Ensemble Models like Random Forest and XG Boost, These models are chosen based on their suitability for detecting and analysing.

The implementation of the machine learning models is done using Python and its libraries for data processing and analysis, such as Django,NumPy, and Pandas.The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology was used to prepare and report the results of this study [5]. PRISMA is a standard method to give a systematic review of existing research

**Step 3: Model Evaluation:**

The final step in this research design is to evaluate the machine learning models based on their performance in detecting and analysing.The models are optimized for performance using techniques such as hyperparameter tuning and cross-validation. The process of identifying eligible articles is depicted in Figure 2. Initially, the queries returned 1796 publications. After that, the publications were screened to eliminate duplicates. There are 473 duplicates that were removed. The abstracts and titles were read in order to examine the techniques and account for the aforementioned inclusion and exclusion criteria, resulting in the removal of 1196 articles and the retention of 127 for a more in-depth examination. Following the full publication review, 14 studies were excluded due to non-English language publications and studies that were unable to get access to the manuscripts. Finally, 113 publications between the year 2001 until 2021 were included in the systematic review. Table 2 summarizes the publications in terms of their type of satellite used, ML techniques involved, water quality parameters extracted and significance or outcomes of studies.

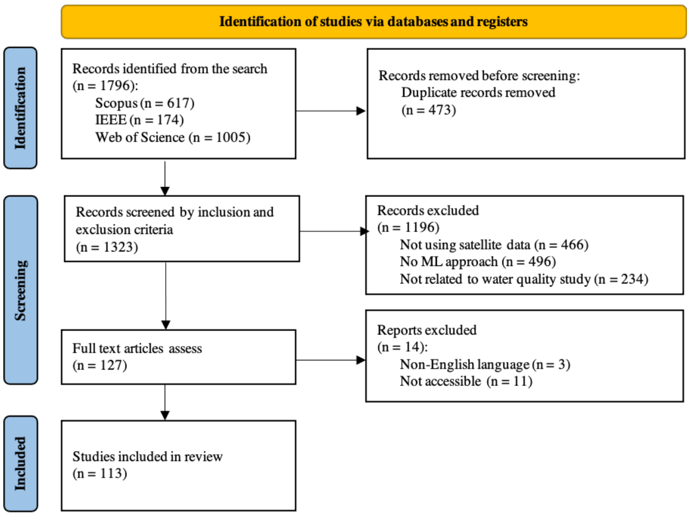


Figure.1 The Proposed system architecture Machine Learning Applicatiom Using Water Quality In Satellite Data

## Data Sets:

The initial crucial stage is to gather data after defining the business problem. It is essential to comprehend the sources of data. The data gathered during this phase is in its raw form, as it may come from various sources and systems, and hence, it is not organized [1].

'name of the customer ', ' village , 'mandal' , 'district' , 'state', 'country', 'temperature', 'humidity', 'pH value' , '

**Python libraries:**

| Library | Purpose |
|---|---|
| Pandas | To read the dataset |
| Django | Setting files and data models |
| NumPy | Used for working with arrays |

Table 1. Python Libraries

## I. CONCLUSIONS:

This systematic review summarized how ML has been applied on satellite data to study water quality issues. The initial search process resulted in 1796 publications, and by refining the search by removing 473 duplicates publication, excluded 1196 non-related topics publications. Through the screening of 127 publications, 113 papers have been selected for data extraction and synthesis. Results also showed that there is a huge variety of ML methods suggested especially on the retrieval of water quality parameters. The most common ML approaches were ANN, SVM, RF, DT, MLP, cubist and GP for monitoring water quality at regional and global scales. According to the systematic analysis, ML can be successfully extended to water quality monitoring, allowing researchers to forecast and learn from natural processes in the environment, as well as assess human impacts on an ecosystem. These initiatives will also aid policymakers and water resource managers in taking proactive actions to prevent the negative consequences of water pollution through restoration projects, as well as ensure that environmental regulatory rules are followed.

## REFERENCES

[1] Diersing N, Keys F and Marine N 2009 Water quality: frequently asked questions Florida Keys Natl. Mar. Sanctuary 8 5–6.

[2] Mohebzadeh H and Lee T 2021 Spatial downscaling of MODIS chlorophyll-a with machine learning techniques over the west coast of the Yellow Sea in South Korea J. Oceanogr. 77 103–22.

[3] Kim H C, Son S, Kim Y H, Khim J S, Nam J, Chang W K, Lee J H, Lee C H and Ryu J 2017 Remote sensing and water quality indicators in the Korean West coast: spatio-temporal structures of MODIS-derived chlorophyll-a and total suspended solids Mar. Pollut. Bull. 121 425–34.

[4] Sagan V, Peterson K T, Maimaitijiang M, Sidike P, Sloan J, Greeling B A, Maalouf S and Adams C 2020 Monitoring inland water quality using remote sensing: potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing Earth-Science Rev. 205 103187.

The Preferred Reporting Items for Systematic Reviews and ML

[6] Guo H, Huang J J, Chen B, Guo X and Singh V P 2021 A machine learning-based strategy for estimating non-optically active water quality parameters using Sentinel-2 imagery Int. J. Remote Sens. 42 1841–66

[7] Tenjo C, Ruiz-verdú A, Van Wittenberghe S, Delegido J and Moreno J 2021 A new algorithm for the retrieval of sun induced chlorophyll fluorescence of water bodies exploiting the detailed spectral shape of water-leaving radiance Remote Sens. 13 1–19.

[8] Du Z, Qi J, Wu S, Zhang F and Liu R 2021 A spatially weighted neural network based water quality assessment method for large-scale coastal areas Environ. Sci. Technol. 55 2553–63.

[9] Bao S, Zhang R, Wang H, Yan H, Chen J and Wang Y 2021 Correction of satellite sea surface salinity products using ensemble learning method IEEE Access 20 1–1.

[10] Xavier Prochaska J, Cornillon P C and Reiman D M 2021 Deep learning of sea surface temperature patterns to identify ocean extremes Remote Sens. 13 1–18.

[11] Maier P M, Keller S and Hinz S 2021 Deep learning with WASI simulation data for estimating chlorophyll a concentration of inland water bodies Remote Sens. 13 1–27. [12] Su H, Lu X, Chen Z, Zhang H and Lu W 2021 Estimating coastal chlorophyll-a concentration from time-series OLCI data based on machine learning Remote Sens. 13 576.

[13] Liu H, Li Q, Bai Y, Yang C, Wang J, Zhou Q, Hu S, Shi T, Liao X and Wu G 2021 Improving satellite retrieval of oceanic particulate organic carbon concentrations using machine learning methods Remote Sens. Environ. 256 112316. [14] Bayati M and Danesh-Yazdi M 2021 Mapping the spatiotemporal variability of salinity in the hypersaline Lake Urmia using Sentinel-2 and Landsat-8 imagery J. Hydrol. 595 126032.

[15] Fan Y, Li W, Chen N, Ahn J H, Park Y J, Kratzer S, Schroeder T, Ishizaka J, Chang R and Stamnes K 2021 OC-SMART: A machine learning based data analysis platform for satellite ocean color sensors Remote Sens. Environ. 253 112236.

[16] Senta A and Šerić L 2021 Remote sensing data driven bathing water quality assessment using Sentinel-3 Indones. J. Electr. Eng. Comput. Sci. 21 1634–47.

[17] Oiry S and Barillé L 2021 Using sentinel-2 satellite imagery to develop microphytobenthosbased water quality indices in estuaries Ecol. Indic. 121 107184.

[18] Cao Z, Ma R, Duan H, Pahlevan N, Melack J, Shen M and Xue K 2020 A machine learning approach to estimate chlorophyll-a from Landsat-8 measurements in inland lakes Remote Sens. Environ. 248 111974.

[19] Hu C, Feng L and Guan Q 2020 A machine learning approach to estimate surface chlorophyll a 11 concentrations in global oceans from satellite measurements IEEE Trans. Geosci. Remote Sens. 59 4590–607. [20] Arnault S, Thiria S, Crépon M and Kaly F 2020 A tropical Atlantic dynamics analysis by

combining machine learning and satellite data Adv. Sp. Res. 68 467-486

[21] Saberioon M, Brom J, Nedbal V, Souček P and CísařP 2020 Chlorophyll-a and total suspended solids retrieval and mapping using Sentinel-2A and machine learning for inland waters Ecol. Indic. 113 106236.

[22] Campbell A M, Racault M F, Goult S and Laurenson A 2020 Cholera risk: A machine learning approach applied to essential climate variables Int. J. Environ. Res. Public Health 17 1–24.

[23] Welch H, Brodie S, Jacox M G, Robinson D, Wilson C, Bograd S J, Oliver M J and Hazen E L 2020 Considerations for transferring an operational dynamic ocean management tool between ocean color products Remote Sens. Environ. 242 111753.

[24] Park J, Kim H C, Bae D and Jo Y H 2020 Data reconstruction for remotely sensed chlorophyll-a concentration in the Ross Sea using ensemble-based machine learning Remote Sens. 12.

[25] Peterson K T, Sagan V and Sloan J J 2020 Deep learning-based water quality estimation and anomaly detection using Landsat-8/Sentinel-2 virtual constellation and cloud computing GIScience Remote Sens. 57 510–25.

[26] Mugo R and Saitoh S I 2020 Ensemble modelling of skipjack tuna (Katsuwonus pelamis) habitats in the western north pacific using satellite remotely sensed data; a comparative analysis using machine-learning models Remote Sens. 12 2591.

[27] Fu Z, Hu L, Chen Z, Zhang F, Shi Z, Hu B, Du Z and Liu R 2020 Estimating spatial and temporal variation in ocean surface pCO2 in the Gulf of Mexico using remote sensing and machine learning techniques Sci. Total Environ. 745 140965.

[28] Sauzède R, Johnson J E, Claustre H, Camps-Valls G and RuescasA B 2020 Estimation of Oceanic Particulate Organic Carbon with Machine Learning ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 5 949–56.

[29] Guan Q, Feng L, Hou X, Schurgers G, Zheng Y and Tang J 2020 Eutrophication changes in fifty large lakes on the Yangtze Plain of China derived from MERIS and OLCI observations Remote Sens. Environ. 246 111890.

[30] DeLuca N M, Zaitchik B F, Guikema S D, Jacobs J M, Davis B J K and Curriero F C 2020 Evaluation of remotely sensed prediction and forecast models for Vibrio parahaemolyticus in the Chesapeake Bay Remote Sens. Environ. 250 112016