Intermediate Project Status Report

Sai Sushma Maddali

San Jose State University

DATA 245: Machine Learning Technologies

Prof. Vishnu S. Pendyala

October 27, 2025

Progress

Significant progress has been achieved in data preparation, cleaning, and analytical setup for the Water Quality prediction project using California Department of Water Resources (DWR) field datasets.

The workflow followed a systematic data-engineering pipeline, starting from raw multi-parameter water-quality data and culminating in a clean, analysis-ready dataset with standardized parameters suitable for WQI computation and sustainability modeling.

The key steps include:

- **Data Import and Understanding**

  The DWR field dataset was explored for its structure and key metadata, such as station_id, station_type, county_name, and parameter, which were identified. The distribution of water quality samples across California counties from 1913 to 2025 is shown in Figure 1. Based on this distribution, only data collected between 2000 and 2025 were used for modeling to ensure consistency and data quality.

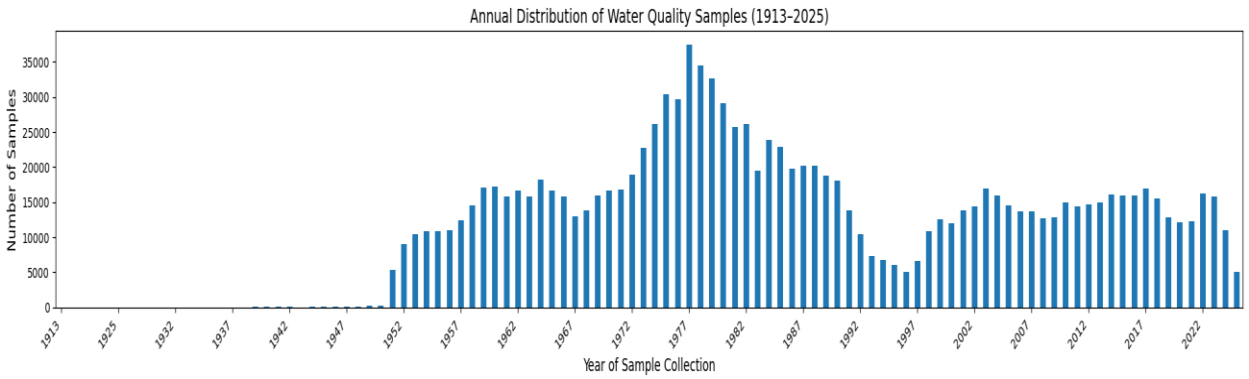


Figure 1: *Distribution of water-quality samples across California counties from 1913 to 2025.*

The sample counts per county are presented in Figure 2, which helped identify regions with limited observations. Counties with sufficient data coverage retrained for model training are shown in Figure 3.

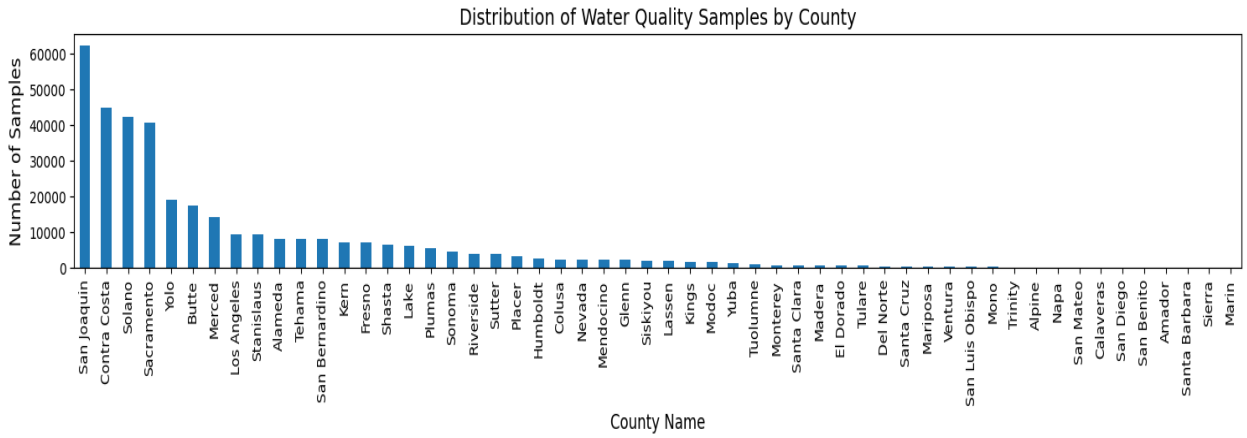


Figure 2: *Number of water-quality samples recorded per county in California.*

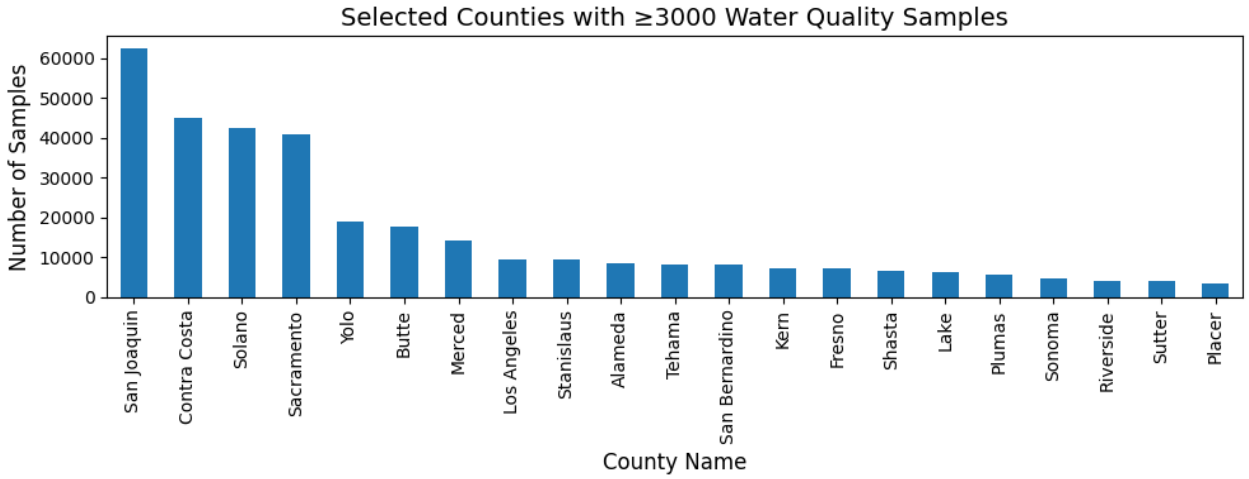


Figure 3: *Counties with sufficient data coverage retained for modeling.*

- **Unit Standardization and Cleaning**

  All parameters were normalized to consistent scientific units (e.g, °C for temperature, µS/cm for conductance, NTU for turbidity, mg/L for dissolved oxygen).

  Mixed units and textual inconsistencies were resolved, ensuring comparability across stations and years.

  The diversity of original unit types is shown in Figure 4, and the post-normalization consistency is illustrated in Figure 5.

	parameter	units_found
0	(Bottom) DissolvedOxygen	[mg/L, % Saturation]
1	(Bottom) SpecificConductance	[uS/cm@25 °C, µmhos/cm@25°C]
2	(Bottom) WaterTemperature	[°C]
3	(Bottom)Chlorophyll Fluorescence	[ug/L of Chl, RFU]
4	(Bottom)Turbidity	[N.T.U., F.N.U.]
5	(Bottom)pH	[pH Units]
6	Carbon Dioxide	[mg/L]
7	Chlorophyll Fluorescence	[RFU, ug/L of Chl]
8	Chlorophyll Volume	[mL]
9	Discharge	[cfs]
10	DissolvedOxygen	[mg/L, % Saturation, %,]
11	ElectricalConductance	[, uS/cm, uS/cm@25 °C]
12	Flow, channel	[cfs, Gallons]
13	Redox Potential	[mV]
14	Secchi Depth	[Meters, Feet, Centimeters,]
15	SoilRedox Potential	[mV]
16	SpecificConductance	[uS/cm@25 °C, µmhos/cm@25°C,]
17	SpecificConductance (EC w/time)	[uS/cm@25 °C]
18	Turbidity	[N.T.U., F.N.U.,]
19	Turbidity (w/time)	[N.T.U.]
20	WaterTemperature	[°C, , °F]
21	WaterTemperature (w/time)	[°C]
22	pH	[pH Units,]
23	pH (w/time)	[pH Units]

Figure 4: *Distribution of unit types across parameters before standardization.*

	parameter	units_found
0	(Bottom) DissolvedOxygen	[mg/L]
1	(Bottom) SpecificConductance	[µS/cm]
2	(Bottom) WaterTemperature	[°C]
3	(Bottom)Chlorophyll Fluorescence	[µg/L]
4	(Bottom)Turbidity	[NTU]
5	(Bottom)pH	[None]
6	Carbon Dioxide	[mg/L]
7	Chlorophyll Fluorescence	[µg/L]
8	Chlorophyll Volume	[µg/L]
9	Discharge	[m³/s]
10	DissolvedOxygen	[mg/L]
11	ElectricalConductance	[µS/cm]
12	Flow, channel	[m³/s]
13	Redox Potential	[mV]
14	Secchi Depth	[m]
15	SoilRedox Potential	[mV]
16	SpecificConductance	[µS/cm]
17	SpecificConductance (EC w/time)	[µS/cm]
18	Turbidity	[NTU]
19	Turbidity (w/time)	[NTU]
20	WaterTemperature	[°C]
21	WaterTemperature (w/time)	[°C]
22	pH	[pH units]
23	pH (w/time)	[pH units]

Figure 5: *Parameter units after standardization to consistent scientific scales.*

- **Outlier Detection and Removal**

  Physically impossible or instrument-error values (e.g., negative temperature, pH > 14, DO > 20 mg/L) were flagged and replaced with NaN.

  A total of 296 outlier records (0.1 %) were removed, ensuring the data reflects realistic environmental conditions.

  A curated dictionary defining valid parameter ranges, based on EPA and WHO standards, is shown in Figure 6.

  Figure 7 highlights parameters flagged as out of range, while Figure 8 displays the cleaned parameter distributions after outlier removal.

```
valid_ranges = {
    'DissolvedOxygen_mg/L': (0, 20),
    '(Bottom) DissolvedOxygen_mg/L': (0, 20),
    'Flow, channel_m³/s': (0, 10000),
    'Discharge_m³/s': (0, 10000),
    'Secchi Depth_m': (0, 50),
    'Turbidity_NTU': (0, 1000),
    '(Bottom)Turbidity_NTU': (0, 1000),
    'Turbidity (w/time)_NTU': (0, 1000),
    'SpecificConductance_µS/cm': (0, 50000),
    'ElectricalConductance_µS/cm': (0, 50000),
    '(Bottom) SpecificConductance_µS/cm': (0, 50000),
    'SpecificConductance (EC w/time)_µS/cm': (0, 50000),
    'WaterTemperature_°C': (-2, 50),
    '(Bottom) WaterTemperature_°C': (-2, 50),
    'WaterTemperature (w/time)_°C': (-2, 50),
    'pH_pH units': (0, 14),
    '(Bottom)pH_None': (0, 14),
    'pH (w/time)_pH units': (0, 14),
    'Chlorophyll Volume_µg/L': (0, 1000),
    'Chlorophyll Fluorescence_µg/L': (0, 1000),
    '(Bottom)Chlorophyll Fluorescence_µg/L': (0, 1000),
    'Redox Potential_mV': (-500, 1000),
    'SoilRedox Potential_mV': (-500, 1000),
    'Carbon Dioxide_mg/L': (0, 200)
}
```

Figure 6: *Reference dictionary of valid parameter value ranges based on EPA and WHO standards.*

```
value_flag
valid          276555
out_of_range      296
Name: count, dtype: int64
           parameter_unit  fdr_result
358   DissolvedOxygen_mg/L  836.000000
764            pH_pH units   16.600000
7066           pH_pH units   24.900000
8160  DissolvedOxygen_mg/L   76.000000
8572           pH_pH units   70.200000
8722   WaterTemperature_°C   -4.222222
9512   WaterTemperature_°C  -13.111111
9565   WaterTemperature_°C   -7.333333
9576   WaterTemperature_°C   -2.833333
9588   WaterTemperature_°C  -12.333333
```

Figure 7: *Parameters flagged as out-of-range prior to data cleaning*

```
Before cleaning:
value_flag
valid          276555
out_of_range      296
Name: count, dtype: int64

Replaced 296 out-of-range values with NaN.

 All parameter values fall within valid environmental ranges.

Summary of cleaned parameter values:
                                   count   min         max  median
parameter_unit
SpecificConductance_µS/cm          52420  0.00  48600.0000  374.85
WaterTemperature_°C                50779 -0.10     50.0000   17.20
pH_pH units                        46299  0.00     14.0000    7.71
DissolvedOxygen_mg/L               44526  0.00     20.0000    8.85
Turbidity_NTU                      37278  0.00   1000.0000    7.80
Secchi Depth_m                     11458  0.00     23.4000    0.80
Chlorophyll Fluorescence_µg/L      10580  0.00    245.0000    2.20
(Bottom) DissolvedOxygen_mg/L       3923  5.62     18.9946    9.79
(Bottom)Chlorophyll Fluorescence_µg/L  3900  0.03     25.7300    1.70
Chlorophyll Volume_µg/L             2540  0.00    550.0000  500.00
```

Figure 8: *Distribution of cleaned parameter values after outlier removal.*

- **Handling Missing and Sparse Parameters**

Missing values were analyzed, and sparse parameters (>80% missing) were excluded. As shown in Figure 9, five core parameters – pH, Dissolved Oxygen, Turbidity, Specific Conductance, and Water Temperature were retained as they covered 70-98% of the data. Missing coordinates were imputed using county-level averages.

	Missing Count	Coverage (%)
DissolvedOxygen_mg/L	10945	79.617111
Secchi Depth_m	42286	21.250722
SpecificConductance_µS/cm	1302	97.575284
Turbidity_NTU	16443	69.378178
WaterTemperature_°C	3011	94.392610
pH_pH units	7428	86.166825

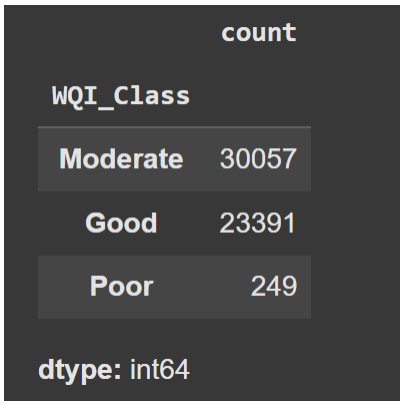
Figure 9: *Coverage percentage of core parameters.*

- **Pivoting and Dataset Consolidation**

  The dataset was transformed from long to wide format, resulting in one row per sampling event with each parameter as a separate feature column while preserving station metadata.

- **Water Quality Index Calculation**

  The cleaned dataset is used to compute WQI using a weighted model based on five key parameters, representing physical, chemical, and biological water quality indicators. The distribution of WQI classes is presented in Figure 10, indicating that the majority of samples fall within the “Good” or “Moderate” categories.



WQI_Class	count
Moderate	30057
Good	23391
Poor	249

dtype: int64

Figure 10: *Distribution of water-quality index (WQI) classes after computation.*

At this stage, the data preparation phase is completed, and the dataset is ready for modelling.

Difficulties

Several technical and methodological challenges were encountered during data preparation and the preprocessing phase. Each issue was analyzed and addressed through custom data engineering strategies.

- **Data Volume and Complexity**

  The DWR datasets exceeded 1.5 GB, which included four tables spanning multiple years and regions. To manage this, google colab is used for handling and processing the data without overloading the system memory. The dataset also contained nested structures and mixed parameter formats across different counties. Each file is explored and filtered to extract relevant parameters required for water quality prediction.

- **Inconsistent Units and Text Variations**

  A major challenge was the inconsistency in parameter measurement units. The same parameter is measured using different units; for example, depth was measured in meters, centimeters, and feet. To ensure meaningful comparison, a conversion dictionary was developed to standardize all measurements to scientifically recognized units, such as NUT for turbidity, °C for temperature, and μS/cm for conductance. After standardization, parameter consistency was verified using descriptive statistics.

- **High Dimensionality and Sparse Parameters**

  The raw dataset contained more than 100 parameters, many of which were not measured consistently across all counties. This high dimensionality made it challenging to identify useful features and risked introducing noise into the model. A parameter coverage analysis was performed to quantify the percentage of missing data per feature. Parameters with more than 80% missing values were excluded to reduce redundancy. Ultimately, five high-coverage parameters were retained for modeling. This streamlined the dataset and improved computational efficiency.

- **Handling Missing Values**

  Missing values were widespread due to irregular measurements across stations and years. Different strategies were applied depending on the type of data.

  Spatial data, such as latitude and longitude, were filled using county-level averages to preserve spatial integrity. For numeric features, a two-level imputation was performed. First, the county-level median value was used when county information is available; otherwise, the global median value was applied. This multi-level imputation framework

minimized data loss and ensured that all key features were complete and ready for model training.

- **Outlier Detection**

  Several extreme readings were identified (e.g, negative temperatures, pH > 14, DO > 20 mg/L). To handle this, parameter-specific validation ranges were established based on EPA and WHO water quality standards. A Python script was used to flag all out-of-range values, which were subsequently replaced with NaN and later handled through imputation.

- **Uneven Temporal Coverage**

  The dataset contained samples ranging from 1913 to 2025, but the early years had very few records compared to modern monitoring programs. A temporal filtering strategy was applied to include only data from 2000-2025, to prevent bias from historical gaps.

  Spatial Imbalance across Counties

  Some counties had tens of thousands of records, while others had fewer than 500 samples. This risked model bias toward over-represented regions. Therefore, Counties with fewer than 3000 records were excluded from the modeling dataset. A county-level median aggregation was performed to ensure spatial representativeness across all retained regions.

Remaining Tasks

The remaining tasks are related to feature engineering, feature selection, model development, and evaluation.

- **Additional Data Preprocessing**

Categorical variables such as station_type and county_name will undergo one-hot encoding, while numeric features will be standardized using StandardScaler.

- **Feature Engineering**

  County-level aggregated information will be added to capture the regional context. Temporal variable will be utilized to include seasonal and long-term trends in water quality.

- **Handling Class Imbalance**

  Since the dataset exhibits class imbalance across water quality categories, the Synthetic Minority Oversampling Technique (SMOTE) will be applied to balance the training data and prevent model bias towards the majority class.

- **Dimensionality Reduction**

  Principal component analysis (PCA) will be used to address the curse of dimensionality.

- **Model Development**

  Multiple machine learning models will be trained and compared, including Support Vector Machines (SVM), Decision Trees, Random Forest, and XGBoost. These algorithms were chosen because of their ability to handle non-linear and complex environmental data effectively.

- **Model Evaluation and Tuning**

  The models will be evaluated using the F1 score, which balances precision and recall, providing a robust metric for imbalanced classification.

  Grid Search will be used for hyperparameter tuning to optimize model parameters and improve predictive performance.

- **Feature Selection**

Post-modeling, the most important features contributing to water quality prediction will be identified using feature importance metrics from Random Forest or XGBoost models. These selected features will enhance model interpretability and guide sustainability insights.

Literature Review

This literature review examines 15 peer-reviewed studies published between 2022 and 2025 that explore the application of machine learning (ML) and related technologies such as deep learning, Internet of Things (IoT), and remote sensing in water quality prediction. The following sections summarize each study's objectives, methodology, and key findings.

Water Quality Prediction with Machine Learning Algorithms

The study by Kumar and Singh (2024) addresses the challenge of accurately forecasting water quality parameters to support sustainable water management systems. The research problem focuses on addressing the inefficiencies of conventional laboratory-based monitoring methods by utilizing data-driven machine learning techniques to analyze large and complex environmental datasets. The authors implemented various machine learning models to process biological, chemical, and physical indicators that influence water quality.

The findings revealed that machine learning significantly reduced the time and cost of monitoring while enhancing prediction accuracy, offering a scalable alternative to manual testing (Kumar & Singh, 2024).

A Review of the application of Machine Learning in Water Quality Evaluation

The review by Zhu et al. (2022) explores how machine learning (ML) is being used to overcome the drawbacks of traditional water quality measurement techniques. The authors utilized algorithms like random forests, support vector machines, neural networks, and ensemble models to predict important water quality indicators - dissolved oxygen, chemical oxygen demand, and nutrient levels.

They compared both supervised and unsupervised models and found that hybrid and deep learning models tend to give more accurate and adaptable results across various environmental conditions. These models were more precise than statistical methods when it came to spotting messy data. The authors also pointed out some challenges, like making models easier to understand, choosing the right data features, and ensuring the input data is of high quality.

The authors highlight machine learning as a valuable tool for achieving Sustainable Development Goal 6, focused on providing clean water and sanitation for everyone (Zhu et al., 2022).

Water Quality Prediction Based on Machine Learning and Comprehensive Weighting Methods

The goal of the research by Wang et al. (2023) is to make water quality predictions more accurate and easier to understand without relying on complicated and data-heavy simulation

models. The researchers utilized real-time data from China's Pearl River Basin and developed a new method for choosing important water quality indicators. They combined Entropy weighting, which measures how much useful information each variable contains, and Pearson correlation, which shows how strongly variables are related, for feature selection. This hybrid method helped the models be easier to interpret and reduce unnecessary complexity.

The authors have trained five different machine learning models, which include SVM, MLP, Random Forest, XGBoost, and LSTM, to predict water quality. The LSTM model outperformed in predicting dissolved oxygen with high accuracy scores (R2 = 0.882, NSE = 0.877). The model was particularly good at recognizing patterns over time and handling noisy data (Wang et al. 2023).

A comparison of Unsupervised and Supervised Machine Learning Algorithms to Predict Water Pollutions

The study by Zamri et al. (2022) explores how different machine learning techniques can help identify pollution levels in Malaysia's Terengganu River. The goal of the research is to find which algorithms work best for spotting pollution hotspots in large, complex water quality datasets. The researchers tested unsupervised models to group data without labeled outcomes and supervised models to make predictions.

They worked with 405 water samples, each describing 27 different water quality features, and classified them into five pollution levels. To improve model performance, they used Standard Scaler for normalization and Principal Component Analysis to reduce the dimensions of the data.

They concluded that after 10-fold cross-validation, Random Forest outperformed all the other models used with an accuracy of 98.78%. Through their experiments, they found out that simpler models like Logistic Regression and LDA also performed well.

The study showed that careful data preparation and tuning can make ML models highly effective for classifying water pollution (Zamri et al., 2022).

Advances in Machine Learning and IoT for Water Quality Monitoring: A Comprehensive Review

The study by Essamlali, Nhaila, and El Khalili (2024) aims to combine Internet of Things (IoT) sensors with ML algorithms to build scalable data systems to collect and analyze environmental data. They had analyzed 141 studies using the structured review method (PRISMA) and focused on how different wireless IoT networks like Zigbee, LoRaWAN work with ML to create real-time, low-cost, and energy-efficient water monitoring systems. They utilized both supervised and unsupervised models to forecast multiple water quality parameters across different environments.

They also highlighted some challenges like making data formats consistent, improving sensor reliability, and securing systems against cyber threats. This tighter integration of IoT with Machine Learning helped cut down on lab testing, energy use, and overall costs, thereby building future-ready water governance systems (Essamlali et al., 2024)

Comparative Analysis of Machine Learning Models for Detecting Water Quality Anomalies in Treatment Plants

The paper by Prabu et al. (2025) focuses on detecting unusual changes in water quality using real-time sensor data. They combined an encoder-decoder neural network with a modified Quality Index (QI). This QI adjusts the importance of water quality parameters based on real-time conditions.

The encoder-decoder model is trained to learn normal sensor behavior and flag unusual patterns that might signal a problem. Before training the model, the data goes through normalization and feature selection to improve model performance. The authors compared different models like ANNs, hybrid models, and transformer-based architectures, using a wide range of evaluation metrics such as Matthews Correlation Coefficient, Critical Success Index, delta-P, accuracy, precision, and recall.

The authors have found a feature called “adaptive QI,” which helps operators make better decisions by turning complex sensor data into a clear, actionable score (Prabu et al., 2025).

A Review of Machine Learning and Internet-of-Things on the Water Quality Assessment: Methods, Applications and Future Trends

The paper by Dharmarathne et al. (2025) explores how the integration between machine learning with the Internet of Things improves the detection, prediction, and management of water pollution events. The authors reviewed a wide range of ML techniques like, regression for predicting pollution patterns, unsupervised anomaly detection for spotting unexpected changes. One major advantage of IoT is its ability to collect high-frequency, location-rich data, which traditional lab-based sampling methods cannot match.

The authors also mentioned the challenges they faced, such as data quality, sensor calibration, model transparency, and deployment costs. They emphasized that successful

water monitoring systems need end-to-end integration from sensors and connectivity to machine learning pipelines and user-friendly interfaces. Most importantly, the paper does not suggest replacing classical water-quality models, but instead, it shows how ML-IoT systems can complement them, offering more flexible and responsive tools for managing freshwater ecosystems (Dharmarathne et al., 2025).

Developing a Novel tool for Assessing the Groundwater Incorporating Water Quality Index and Machine Learning Approach

The study by Sajib et al. (2023) explores a smarter way to assess groundwater risks by combining machine learning with GIS-based multi-criteria decision-making (MCDM). The authors built a framework that layers together different types of spatial data, hydrogeology, land use, and climate conditions, and further used MCDM to assign weights based on their importance. These weighted features are fed to machine learning models such as decision trees and support vector machines to predict which areas are the most vulnerable to groundwater issues.

One of the key strengths of this approach is that it blends spatial context with ML's ability to handle uncertainty. To test the tool, they applied it to a real-world case study and compared the model's predictions with actual groundwater quality and supply data. The results turned out to be impressive as the hybrid model was very accurate in flagging high-risk zones that needed attention (Sajib et al., 2023).

Machine Learning Approach for Water Quality Predictions Based on Multispectral Satellite Imageries

The study by Anand, Oinam, and Wieprecht (2024) explores how combining machine learning with satellite imagery can make water quality monitoring more efficient and affordable, especially in places where field sampling is difficult or costly. Using data from two satellites, Sentinel-2 and ResourceSat-2(LISS-IV), the researchers-built models such as Support Vector Machines, Random Forest, and Multiple Linear Regression to estimate key water quality indicators. The satellite-based maps revealed clear seasonal and regional shifts in water quality, reflecting the dynamic ecosystem of the water bodies.

This research shows that machine learning with remote sensing is a powerful tool for real-time, large-scale water monitoring. It helps track pollution sources and seasonal changes without needing constant fieldwork (Anand et al., 2024).

Evaluation of Water Quality Indexes with Novel Machine Learning and Shapley Additive ExPlanation (SHAP) Approaches

The paper by Aldreese et al. (2024) explores how interpretable machine learning can improve the way water quality is assessed. The authors tested Gene Expression Programming models, Deep Neural Networks, and Optimizable Gaussian Process Regressors to see which models, when paired with explainability tools, offer the most accurate and transparent predictions of water-quality indexes.

They utilized SHAP (Shapley Additive exPlanations) to find the most influential input features. SHAP analysis revealed that bicarbonate, calcium, sulphate, sodium, and magnesium were the most important factors driving electrical conductivity and TDS level, while pH and chloride had less impact. The study shows how ML tools can complement

traditional hydraulic and biogeochemical models by providing high-resolution empirical insights (Aldreese et al., 2024).

Using Ensemble Machine Learning to Predict and Understand Spatiotemporal Water Quality Variations Across Diverse Watersheds in Coastal Urbanized Areas

This study by Xiao et al. (2025) addresses the challenge of predicting water quality changes across multiple urban coastal watersheds, where geography, climate, and human activity all interact. The authors introduced a model called the Ensemble Across-Watersheds Model, which blends five machine learning models into a stacked ensemble for better accuracy. They trained the model on a massive weekly water quality dataset from 432 sites across 12 watersheds. Using SHAP, they identified that tree cover over 55%, building and road density, and proximity to the sea (less than 10 km) as key geographic drivers.

Spatial analysis showed that highly urbanized and coastal areas had the worst water quality, while temporal trends revealed degradation during hot and rainy months. They stated that just 20-40% of samples from urban and coastal zones during extreme weather accounted for most of the model's insights. Thus, it is concluded that targeted sampling in these areas could cut monitoring effort by up to 80% without losing accuracy (Xiao et al., 2025).

Beyond Tides and Time: Machine Learning's Triumph in Water Quality Forecasting

The study by Li et al. (2023) explores how machine learning models can accurately forecast pH levels in surface water systems using routine sensor data, without relying on complex spatial-temporal modeling. The authors tested five ML approaches on daily water quality data. They mentioned that LightGBM delivered the best performance, and tree-based models

consistently outperformed both linear regression and MLP, which was sensitive to feature scaling.

The authors stated that the ML models outperformed a benchmark spatial-temporal model (SDAL-II), showing that with strong feature engineering and tuning, “black-box” models can exceed traditional approaches. The study also highlights the potential for ML pipelines to be deployed by non-experts, expanding access to real-time monitoring in resource-limited regions (Li et al., 2023).

Water Quality Prediction Method Based on a Combined Machine Learning Model: A Case Study of the Daling River Basin

The study by Liu and Wang (2025) introduces a hybrid machine-learning framework, Enhanced Long Short-Term Memory with Back Propagation (ELSTM-EBP), to improve the prediction of total nitrogen levels as a key indicator of aquatic pollution. The model addresses missing data and nonlinearity by combining spatiotemporal interpolation, weighted feature selection, and time-series decomposition. The model identified water temperature, dissolved oxygen, turbidity, and precipitation as dominant predictors. The researchers modeled each of the water-quality components separately before merging outputs for the final forecast.

Seasonal and spatial analysis revealed a “U-shaped” total nitrogen trend and downstream concentration peaks. SHAP analysis confirmed temperature and dissolved oxygen as key drivers, with rainfall amplifying total nitrogen during flood seasons (Liu & Wang, 2025).

A Machine Learning Predictive Model to Detect Water Quality and Pollution

The study by Xu et al. (2022) presents a machine learning framework for classifying water and sediment pollution levels in marine environments, even when a large proportion of data (about 57%) is missing. The authors evaluated various imputation strategies and trained classifiers using support vector machines and neural networks on samples labeled across four pollution levels. The authors identified a strong link between sediment metal concentrations and water pollution.

Their best model achieved around 75% accuracy, demonstrating that automated pollution screening is feasible under data-scarce conditions (Xu et al., 2022).

Forecasting Water Pollution Index Using Machine Learning and Multi-Parameter Water Quality Data

The study by Kolosov et al. (2024) explores how machine learning can improve forecasting of the Water Pollution Index (WPI) using multi-parameter water quality data. The researchers have analyzed 20 physicochemical indicators collected every 10 days from two monitoring sites and trained linear regression, random forest, and XGBoost models on raw, standardized, and polynomial-transformed data.

They stated that the Random Forest model on standardized inputs delivered the most accurate predictions, highlighting the importance of data preprocessing. By linking WPI to multivariate predictions, the study provides a practical framework for operational monitoring and early detection of pollution trends (Kolosov et al., 2024).

Summary

Across the 15 reviewed studies, machine learning has consistently demonstrated its potential to improve water quality monitoring, forecasting, and decision making. Techniques such as LSTM and SHAP-based interpretability have proven particularly effective for managing complex environmental data. Integrating ML with IoT and remote sensing further enhances scalability.

References

Kumar, R., & Singh, A. (2024). *Water quality prediction with machine learning algorithms. EPRA International Journal of Multidisciplinary Research (IJMR), 10(4),* 45-53. https://doi.org/10.36713/epra16318

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B., & Ye, L. (2022). *A review of the application of machine learning in water quality evaluation. Eco-Environment & health, 1(2),* 107-116. https://doi.org/10.1016/j.eehl.2022.06.001

Wang, X., Li, Y., Qiao, Q., Tavares, A., & Liang, Y. (2023*). Water quality prediction based on machine learning and comprehensive weighting methods. Entropy, 25(8),* 1186. https://doi.org/10.3390/e25081186

Zamri, N., Pairan, M. A., Wan Azman, W. N. A., Abas, S. S., Abdullah, L., Naim, S., Tarmudi, Z., & Gao, M. (2022). *A comparison of unsupervised and supervised machine learning algorithms to predict water pollutions. Procedia Computer Science, 217,* 1816-1826. https://doi.org/10.1016/j.procs.2022.08.021

Essamlali, I., Nhaila, H., & El Khalili, M. (2024). *Advances in Machine Learning and IoT for Water Quality Monitoring: A Comprehensive Review. Heliyon, 10(6),* e27920. https://doi.org/10.1016/j.heliyon.2024.e27920

Prabu, P., Alluhaidan, A.S., Aziz, R., & Basheer, S. (2025). *Comparative analysis of machine learning models for detecting water quality anomalies in treatment plants. Scientific Reports, 15,* Article 30453. https://www.nature.com/articles/s41598-025-15517-4

Dharmarathne, G., Abekoon, A. M. S. R., Bogahawattha, M., Alawatugoda, J., & Meddage, D, P. P. (2025). *A review of machine learning and internet-of-things on the water quality assessment: Methods, applications and future trends. Engineering Reports, 7(5),* 105182. https://doi.org/10.1016/j.rineng.2025.105182

Sajib, A. M., Diganta, M. T. M., Rahman, A., Dabrowski, T., Olbert, A. I., & Uddin, M. G. (2023). *Developing a novel tool for assessing the groundwater incorporating water quality index and machine learning approach. Journal of Groundwater for Sustainable Development, 23,* 101049. https://doi.org/10.1016/j.gsd.2023.101049

Anand, V., Oinam, B., & Wieprecht, S. (2024). *Machine learning approach for water quality predictions based on multispectral satellite imageries. Ecological Informatics, 84,* 102868. https://doi.org/10.1016/j.ecoinf.2024.102868

Aldreese, A., Khan, M., Taha, A.T.B., & Ali, M. (2024). *Evaluation of water quality indexes with novel machine learning and Shapley Additive ExPlanation (SHAP) approaches. Journal of Water Process Engineering, 75,* 104789. https://doi.org/10.1016/j.jwpe.2024.104789

Xiao, F., Zhang, R., Jian, Z., Liu, W., Sun, T., Pang, W., Han, L., & Qin, H. (2025). *Using ensemble machine learning to predict and understand spatiotemporal water quality*

variations across diverse watersheds in coastal urbanized areas. *Ecological Indicators*, 178, 113976. https://doi.org/10.1016/j.ecolind.2025.113976

Li, Y., Mao, S., Yuan, Y., Wang, Z., Kang, Y., & Yao, Y. (2023). *Beyond tides and time: Machine learning's triumph in water quality forecasting. arXiv preprint arXiv:2309.16951*. https://arxiv.org/abs/2309.16951

Liu, Y., & Wang, Y. (2025). Water quality prediction method based on a combined machine learning model: A case study of the Daling River Basin. *Journal of Contaminant Hydrology*, *276*, 104725. https://doi.org/10.1016/j.jconhyd.2025.104725

Xu, X., Lai, T., Jahan, S., Farid, F., & Bello, A. (2022*). A machine learning predictive model to detect water quality and pollution. Future Internet, 14(11)*, 324. https://doi.org/10.3390/fi14110324

Kolosobv, D., Ivanov, P., & Makarova, E. (2024). *Forecasting water pollution index using machine learning and multi-parameter water quality data. In Preceedings of the 5th International Workshop on Data Science for Environmental and Ecological Sustainability (DSEES 2024)* (Vol. 3974, pp. 1-6). CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3974/short01.pdf

Appendix A

Project Repository

The data preprocessing scripts used for this project is available in the public GitHub repository:

https://github.com/sai-sushma-maddali/water-quality-prediction/blob/a6b7929bfa5c229a4c28365cfde1c71deaa9c79e/01%20notebooks/water_quality_analysis_eda.ipynb