Leveraging Machine Learning for Sustainable Water Quality Modeling

Sai Sushma Maddali
SJSU ID: 018195775
Email: saisushma.maddali@sjsu.edu
Department of Applied Data Science
San José State University

Abstract—Clean and safe water is fundamental to public health and sustainable resource management. This study presents a machine learning framework for the automated classification of water quality using field observations from the California Department of Water Resources (DWR). The system predicts water-station conditions as Good, Moderate, Poor, or Very Poor based on key physicochemical parameters, including dissolved oxygen, pH, turbidity, conductivity, and temperature. The workflow integrates data preprocessing, feature engineering, and model evaluation within a reproducible pipeline. Several supervised learning algorithms – Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost were trained using a stratified dataset representing statewide water stations. Model performance was evaluated using accuracy, precision, recall, and macro F1-score. Among the tested algorithms, XGBoost achieved the highest predictive accuracy (98.6%) with balanced recall across all classes. The approach demonstrates how open environmental data and interpretable ML models can enhance real-time water-quality assessment, supporting UN Sustainable Development Goal 6 (Clean Water and Sanitation) through scalable, data-driven environmental monitoring.

Index Terms—Machine Learning (ML), Internet of Things (IoT), Support Vector Machine (SVM), Random Forest (RF), Shapely Additive Explanations (SHAP), Sustainable Development Goals (SDG 6), California Department of Water Resources (DWR), United States Environmental Protection Agency (US EPA), World Health Organization (WHO)

1 Introduction

Access to clean and safe water is fundamental for human health, environmental protection, and sustainable development. However, the growing impact of industrialization, agricultural runoff, and climate change has made continuous water-quality monitoring increasingly complex and resource-intensive. Traditional laboratory testing, while accurate, is slow, expensive, and often unable to provide real-time insights necessary for early contamination detection. As a result, there is a critical need for scalable, data-driven solutions that can automate the assessment of water quality and support evidence-based decision-making.

The California Department of Water Resources (DWR) maintains extensive environmental monitoring programs, collecting field and laboratory data from thousands of surface and groundwater stations across the state. Although these data are publicly available through data.ca.gov, their high dimensionality, inconsistent units, missing values, and temporal irregularities make them difficult to use directly for large-scale analysis. The challenge lies in transforming these heterogeneous datasets into a standardized, machine-learning-ready format that can accurately classify and predict water-quality conditions.

This project addresses these challenges by developing a machine-learning pipeline that predicts water quality categories – Good, Moderate, Poor, or Very Poor based on field-measured physicochemical parameters such as dissolved oxygen, pH, turbidity, conductivity, and temperature. The workflow integrates robust data preprocessing, feature engineering, and model training using algorithms including Logistic Regression, Support Vector Machines (SVM), Random Forests, and XGBoost. To ensure interpretability, SHAP (Shapely Additive Explanations) is employed to quantify each parameter's contribution to model predictions, enhancing transparency and trust in the results.

The study not only demonstrates how open environmental data can be leveraged for automated water-quality assessment but also aligns with United Nations Sustainable Development Goal (SDG) 6 – Clean Water and Sanitation. By combining reproducible open-source methods, explainable machine learning, and sustainable data practices, this work contributes to the broader goal of developing intelligent environmental monitoring systems that support timely intervention, safeguard public health, and promote long-term water-resource sustainability.

2 Literature Review

Recent research between 2022 and 2025 demonstrates a surge in the application of machine learning (ML), deep learning, and Internet of Things (IoT) technologies for water quality prediction, monitoring, and forecasting. This section reviews 15 key peer-reviewed studies organized into four thematic categories:

1) General ML models for water-quality prediction,
2) Hybrid and weighting frameworks,
3) IoT and remote-sensing integrations, and
4) Explainable and ensemble approaches.

Department of Applied Data Science, San José State University

2.1 A. Machine Learning Models for Water-Quality Prediction

Kumar and Singh [1] applied supervised ML algorithms to forecast biological and physicochemical parameters, reporting reduced monitoring cost and improved accuracy compared to manual laboratory analysis. Zamri et al. [4] compared supervised and unsupervised algorithms on Malaysian river data and achieved 98.78% accuracy using Random Forest after normalization and PCA. Xu et al. [14] addressed data scarcity in marine environments by evaluating imputation and classification strategies; their SVM-based model achieved 75% accuracy despite 57% missing data. Together, these studies established that robust preprocessing and classical ML models can effectively classify water-quality states.

2.2 Hybrid, Weighting, and Temporal Frameworks

To enhance interpretability and capture nonlinear temporal dynamics, several hybrid approaches were developed. Wan et al. [3] introduced entropy–correlation weighting for feature selection and demonstrated that an LSTM model achieved $R^2 = 0.882$ for dissolved-oxygen prediction. Li et al. [12] compared tree-based models on daily sensor data and found that LightGBM outperformed a spatial-temporal baseline. Liu and Wang [13] proposed an enhanced LSTM–BP hybrid (ELSTM-EBP) integrating interpolation and weighted feature selection, accurately modeling total nitrogen trends. These studies confirm that integrating statistical weighting with time-series architectures improves long-term water-quality forecasting.

2.3 IoT, Remote-Sensing, and GIS Integrations

IoT and geospatial data have greatly expanded the scalability of water-quality assessment. Essamlali et al. [5] reviewed 141 studies combining wireless IoT protocols (Zigbee, LoRaWAN) with ML for low-power, real-time sensing. Dharmarathne et al. [7] emphasized challenges in data quality, calibration, and security while advocating for fully integrated IoT-ML systems. Anand et al. [9] utilized Sentinel-2 and ResourceSat-2 imagery with SVM, RF, and MLR models to map seasonal and spatial variations, proving remote sensing as an effective alternative to field sampling. Sajib et al. [8] fused ML with GIS-based multi-criteria decision making for groundwater risk mapping, achieving accurate identification of high-risk zones. These integrations demonstrate ML's role in sustainable, large-scale monitoring infrastructures.

2.4 Explainability, Ensemble, and Anomaly-Detection Approaches

Interpretability and adaptive modeling have become essential for operational deployment. Aldreese et al. [10] coupled SHAP analysis with Gene-Expression Programming and Gaussian process models, revealing bicarbonate, calcium, and sulphate as key conductivity drivers. Xiao et al. [11] developed a cross-watershed ensemble combining five learners; SHAP interpretation linked tree cover and coastal proximity to water-quality degradation, reducing sampling effort by 40%. Prabu et al. [6] proposed an encoder-decoder neural model with an adaptive Quality Index to detect anomalies in treatment plant sensors. Kolosov et al. [15] used Random Forest on standardized inputs to forecast the Water Pollution Index, confirming the importance of data preprocessing. Zhu et al. [2] provided a comprehensive review emphasizing that hybrid and ensemble models outperform traditional statistical techniques and directly support Sustainable Development Goal 6.

2.5 Summary and Research Gap

Across the 15 reviewed studies, ML consistently enhanced the accuracy, scalability, and cost efficiency of water-quality prediction. Temporal models (LSTM), ensemble methods, and explainability tools (SHAP) enabled transparent and actionable insights. Integrating IoT and remote sensing further improved spatial coverage and real-time monitoring. However, reproducibility and feature-standardization remain limited across regions. This project addresses these gaps by developing a unified, station-wise preprocessing and classification pipeline to support sustainable water-management systems aligned with SDG 6.

3 Dataset Description

This study utilizes the Field Results dataset provided by the California Department of Water Resources (DWR) through the official data.ca.gov portal. The dataset contains direct field measurements collected at multiple water-quality monitoring stations across California. It forms part of the DWR's long-term environmental monitoring program that adheres to standardized field sampling protocols for both surface and groundwater systems.

3.1 Data Characteristics

The original dataset comprises 1,217,203 records with 22 attributes, covering 29,271 monitoring stations and approximately 65 distinct physicochemical parameters measured across the state. Each record in the dataset corresponds to a unique sampling event, defined by a combination of station identifier, date, depth, and parameter name. The dataset includes on-site physicochemical measurements such as:

1) Dissolved Oxygen (mg/L)
2) pH
3) Turbidity (NTU)
4) Specific Conductance (µS/cm)
5) Water Temperature (°C)
6) Secchi Depth (m)
7) Chlorophyll Fluorescence (RFU), among others

Figure 1 shows the frequency distribution of the leading physicochemical parameters in the dataset, illustrating the varying measurement density across different water quality indicators.

Additional metadata fields describe station identifiers and geographic coordinates (latitude, longitude), station type (surface or groundwater), county name, sampling date, reporting limit, and measurement units.

The dataset covers observations from 1913 to 2025, as shown in Figure 2. However, only data collected after 2000 were used for analysis to ensure uniformity in measurement

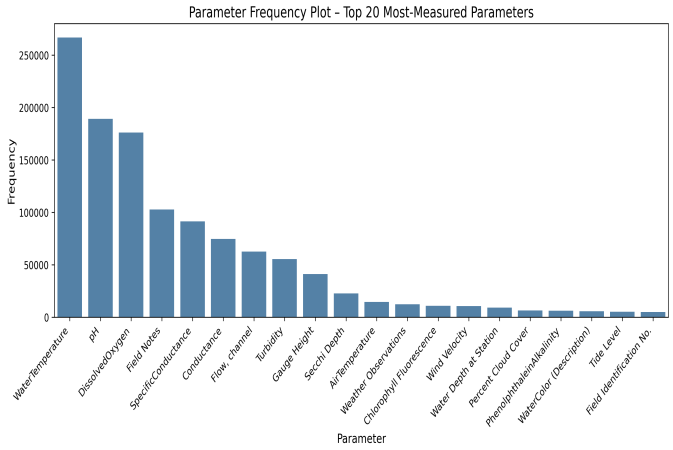

Fig. 1: Frequency distribution of leading physicochemical parameters.

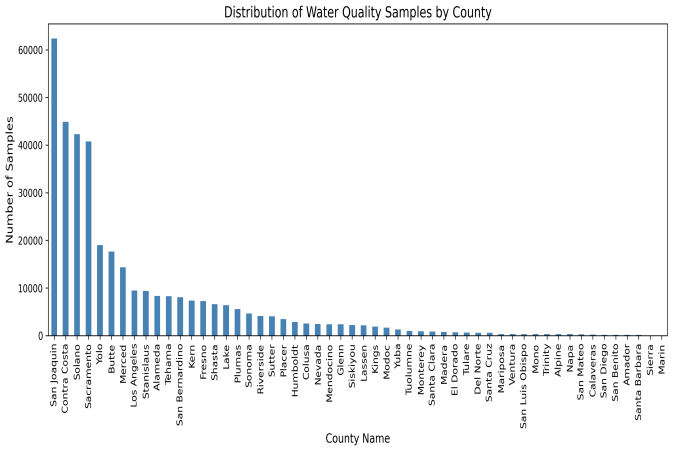

Fig. 3: Distribution of water quality samples across California counties.

methods and instrumentation. This temporal filtering addresses the significant increase in sampling frequency and standardization in modern monitoring practices.

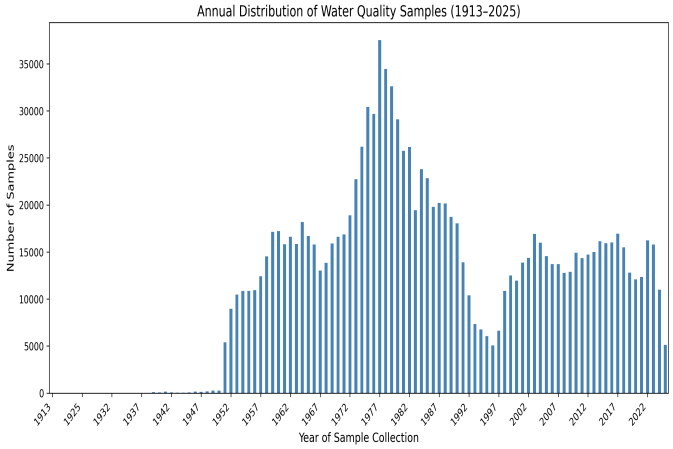

Fig. 2: Temporal distribution of water quality samples from 1913 to 2025.

The spatial distribution of samples across California counties is illustrated in Figure 3. The dataset shows considerable geographic variation, with certain counties having substantially more monitoring stations and sampling events than others. This spatial imbalance was addressed during preprocessing by applying geographic sampling thresholds to ensure adequate representation.

The subset used for modeling consists of approximately 50,000 valid records after filtering. All data are supplied in CSV format, which ensures interoperability with data science tools and facilitates reproducibility.

4 Data Preprocessing

Comprehensive preprocessing was applied to the DWR Field Results dataset to transform raw measurements into a modeling-ready form as shown in in Figure 4. The workflow included unit standardization, outlier removal, missing-value imputation, dataset reshaping, and Water Quality Index (WQI) computation.

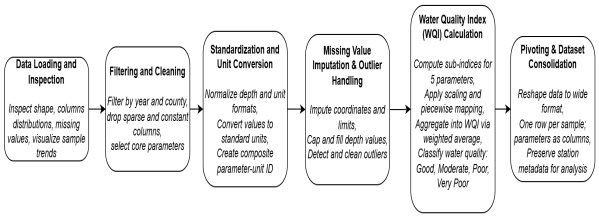

Fig. 4: Workflow illustrating the data preprocessing steps applied prior to model training.

4.1 Temporal and Geographic Scope Filtering

The process began with systematic data filtering to establish a consistent and reliable dataset, addressing temporal and geographic biases present in the raw data. This step provided a statistically sound foundation for model development, focused on high-quality samples.

To address methodological inconsistencies in historical records, the dataset was filtered temporally. Only samples collected between 2000 and 2025 were retained, focusing the analysis on modern and digitally consistent monitoring data.

To ensure adequate geographic representation and prevent model bias toward sparsely sampled regions, the analysis was restricted to 15 counties. These counties were selected based on robust sampling efforts, each contributing at least 3,000 observations and collectively representing the majority of field samples shown in in Figure 5.

Before complex transformations, a completeness check was performed on all features. Any parameter with more than 50% missing values was excluded from the dataset. This step eliminated features that were too sparse to provide reliable information, thereby reducing high-dimensional

noise in the data.

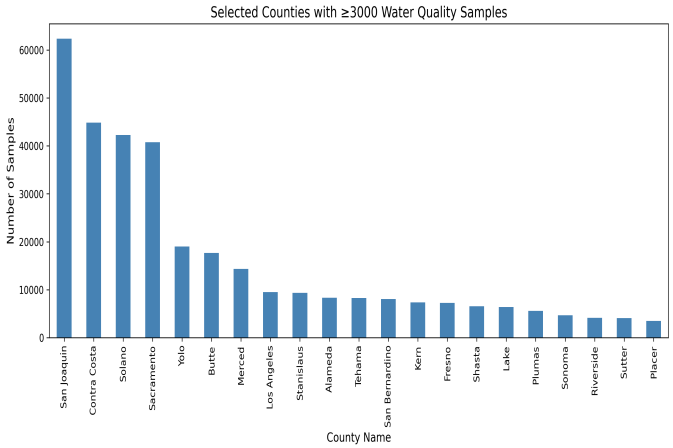

Fig. 5: Geographic distribution of samples by retained counties.

4.2 Unit Standardization and Cleaning

The raw field data contained multiple unit representations for identical physicochemical parameters, depending on measurement instruments and data-entry conventions across monitoring stations. For example, temperature was recorded in both degrees Fahrenheit and degrees Celsius, depth measurements appeared in meters, centimeters, and feet, and electrical conductivity was reported using symbols such as *µS/cm*, *uS/cm*, or *mS/cm*. These inconsistencies made direct comparison and aggregation across counties impractical.

To ensure consistency, all physicochemical parameters were normalized to standard scientific units–°C for temperature, µS/cm for conductance, NTU for turbidity, and mg/L for dissolved oxygen. A custom mapping dictionary was implemented to identify and convert nonstandard or legacy unit formats into their standardized equivalents using predefined conversion factors. For instance, temperature values recorded in °F were converted to °C using $(T_C = (T_F - 32) \times 5/9)$, and conductance values expressed in mS/cm were multiplied by 1,000 to obtain µS/cm. As shown in Table 1 and Table 2, illustrate the diversity of unit types before cleaning and the resulting uniform representation after normalization. The standardization process ensured compatibility across 29,271 monitoring stations and enabled accurate computation of derived metrics such as sub-indices (q_i) and the Water Quality Index (WQI).

4.3 Outlier Detection and Removal

Physically implausible readings—such as negative water temperatures, pH values exceeding 14, or dissolved oxygen concentrations above 20 mg/L—were identified and replaced with missing values. These anomalies typically result from sensor calibration issues, transcription errors, or equipment malfunction during field measurements. Approximately 296 records (0.1% of the dataset) were removed to maintain environmental realism and prevent model bias.

To ensure the data represented scientifically valid observations, acceptable parameter ranges were established using guidelines from the U.S. Environmental Protection Agency (EPA), the World Health Organization (WHO), and the American Public Health Association (APHA). These thresholds were based on the physical and ecological characteristics of natural aquatic systems. For example, dissolved oxygen levels in freshwater rarely exceed 20 mg/L, and pH values in natural water bodies typically range from 6.5 to 8.5.

Table 3 summarizes the valid environmental limits applied consistently across all monitoring stations. Values outside these scientifically supported ranges were considered unrealistic and excluded from analysis. This quality control procedure ensured that the final dataset used for Water Quality Index (WQI) computation and machine learning model training adhered to established environmental standards, enhancing both data integrity and model reliability.

TABLE 1: Units Found Before Standardization

Parameter	Units Found
(Bottom) DissolvedOxygen	mg/L, % Saturation
(Bottom) SpecificConductance	uS/cm@25 °C, µmhos/cm@25°C
(Bottom) WaterTemperature	°C
(Bottom) Chlorophyll Fluorescence	ug/L of Chl, RFU
(Bottom) Turbidity	N.T.U., F.N.U.
(Bottom) pH	pH Units
Carbon Dioxide	mg/L
Chlorophyll Fluorescence	RFU, ug/L of Chl
Chlorophyll Volume	mL
Discharge	cfs
DissolvedOxygen	mg/L, % Saturation, %
ElectricalConductance	uS/cm, uS/cm@25 °C
Flow, channel	cfs, Gallons
Redox Potential	mV
Secchi Depth	Meters, Feet, Centimeters
SoilRedox Potential	mV
SpecificConductance	uS/cm@25 °C, µmhos/cm@25°C
SpecificConductance (EC w/time)	uS/cm@25 °C
Turbidity	N.T.U., F.N.U.
Turbidity (w/time)	N.T.U.
WaterTemperature	°C, , °F
WaterTemperature (w/time)	°C
pH	pH Units
pH (w/time)	pH Units

TABLE 2: Units Found After Standardization

Parameter	Units Found
(Bottom) DissolvedOxygen	mg/L
(Bottom) SpecificConductance	µS/cm
(Bottom) WaterTemperature	°C
(Bottom) Chlorophyll Fluorescence	µg/L
(Bottom) Turbidity	NTU
(Bottom) pH	pH units
Carbon Dioxide	mg/L
Chlorophyll Fluorescence	µg/L
Chlorophyll Volume	µg/L
Discharge	m^3/s
DissolvedOxygen	mg/L
ElectricalConductance	µS/cm
Flow, channel	m^3/s
Redox Potential	mV
Secchi Depth	m
SoilRedox Potential	mV
SpecificConductance	µS/cm
SpecificConductance (EC w/time)	µS/cm
Turbidity	NTU
Turbidity (w/time)	NTU
WaterTemperature	°C
WaterTemperature (w/time)	°C
pH	pH units
pH (w/time)	pH units

4.4 Handling Missing and Sparse Parameters

A structured imputation strategy was used to handle missing data while keeping as many samples as possible for model training. The approach differed based on the type of information being imputed.

TABLE 3: Valid Environmental Ranges for Core Water-Quality Parameters

Parameter	Valid Range	Reference Source
Dissolved Oxygen (mg/L)	0–20	WHO [19], EPA [18]
Water Temperature (°C)	-2–50	EPA [18], WHO [19]
pH (pH units)	0–14 (optimal 6.5–8.5)	EPA [17], WHO [19]
Specific Conductance (µS/cm)	0–50,000	USGS [20]
Turbidity (NTU)	0–1000	WHO [19]
Secchi Depth (m)	0–50	OECD [21]
Chlorophyll Fluorescence (µg/L)	0–1000	UNESCO [22]
Redox Potential (mV)	-500–1000	APHA [23]
Carbon Dioxide (mg/L)	0–200	Wetzel [24]

Parameter Selection Based on Coverage: All water quality parameters were evaluated for data completeness. As illustrated in the missing values heatmap (Fig. 6), the extent of missing data varied significantly across parameters. Parameters with more than 80% missing values were removed because they were too incomplete to contribute reliable information. This process identified five core parameters with good coverage (70-98%): pH, Dissolved Oxygen, Turbidity, Specific Conductance, and Water Temperature. These parameters are key indicators of water quality and have sufficient data for meaningful analysis.

Geographic Location Imputation: Accurate location data is important for understanding regional patterns in water quality. When station coordinates (latitude and longitude) were missing, they were estimated using the average location of other stations in the same county. This county-level approach maintained geographic context and provided reasonable location estimates based on nearby monitoring sites, rather than using a statewide average that could misrepresent the station's actual region.

Water Quality Parameter Imputation: After all filtering and quality checks, some missing values remained in the five core parameters. These gaps were filled using the median value of each parameter. The median was chosen instead of the mean because it's less affected by extreme values or outliers that sometimes occur in environmental data. This ensured that the imputed values represented typical conditions and maintained realistic distributions before calculating the Water Quality Index and training the models.

4.5 Pivoting and Dataset Consolidation

Transforming the data from its original long format to a wide format suitable for machine learning was one of the most critical preprocessing steps. This restructuring consolidated multiple measurements from a single sampling event into one comprehensive row.

The Challenge of the Long Format

The raw data was stored in long format, where each record represented a single measurement:

- A single sampling event (identified by `sample_code`) often spanned multiple rows
- Each row contained the same station metadata (latitude, date, county) but held only one parameter and its value (e.g., a pH reading or dissolved oxygen measurement)

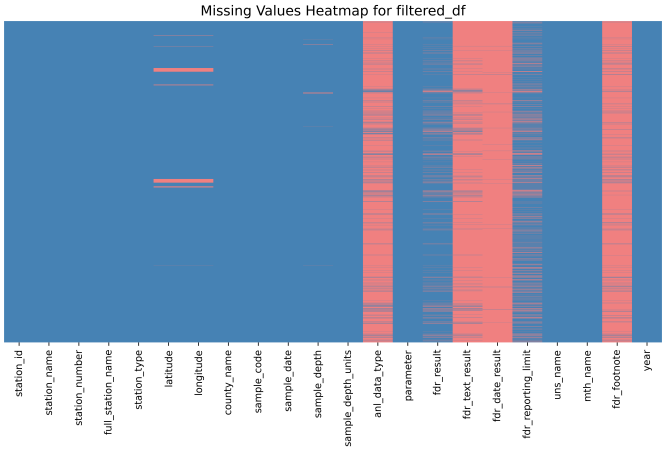

Fig. 6: Heatmap showing the distribution of missing values across features in the dataset.

While this structure works well for data collection, it's incompatible with machine learning models, which require all predictor variables (features) and the target variable to be aligned in a single row.

The Transformation: Pivoting to Wide Format

The dataset was reshaped from long to wide format using a pivot table operation:

Row Identity: Each unique sampling event (`sample_code`) became a single row, representing a distinct moment in time and space.

Column Creation: Standardized parameter names (e.g., `DissolvedOxygen_mg/L`, `pH_pH units`) were converted from row values into separate feature columns.

Aggregation: When a parameter was measured multiple times during the same event (rare but possible), the mean of those measurements was calculated to ensure a single representative value.

Metadata Preservation: Important station information (latitude, longitude, sample date, county) was retained and linked to the corresponding sampling event in the wide format.

Benefits of the Wide Format

This restructuring was essential for three main reasons:

Machine Learning Compatibility: The wide format is the standard input structure for supervised learning algorithms. Models can now simultaneously consider all five core water quality parameters to predict the WQI class.

Easier Analysis and Correlation: With parameters aligned as columns, direct analysis becomes straightforward. Correlation coefficients between parameters (e.g., pH vs. dissolved oxygen) can be easily calculated, providing insights into water chemistry. Feature distributions and multivariate relationships can also be visualized directly.

Target Variable Integration: The transformation creates a unique reference point for each sampling event, allowing the calculated WQI score and WQI class to be added as the target column for that specific sample, completing the dataset for model training.

4.6 Water-Quality Index Calculation

Sub-indices (q_i values) were calculated for the five retained parameters using standard water-quality formulas:

$$\begin{aligned} q_{\text{DO}} &= \min\left(100, \frac{\text{DO}}{14} \times 100\right) \\ q_{\text{pH}} &= \begin{cases} 100 & \text{if } 6.5 \leq \text{pH} \leq 8.5 \\ 100 - 10|\text{pH} - 7.5| & \text{otherwise} \end{cases} \\ q_{\text{Cond}} &= \max\left(0, 100 - \frac{\text{Cond}}{1500} \times 100\right) \\ q_{\text{Turb}} &= \max\left(0, 100 - \frac{\text{Turb}}{100} \times 100\right) \\ q_{\text{Temp}} &= \max\left(0, 100 - |\text{Temp} - 20| \times 5\right) \end{aligned} \tag{1}$$

A weighted average of the sub-indices produced the Water Quality Index (WQI):

$$\text{WQI} = \frac{\sum w_i q_i}{\sum w_i} \tag{2}$$

Using Equations (1) and (2), the WQI was computed for each sample in the dataset.

The assigned weights are $w_{\text{DO}} = 0.3$, $w_{\text{pH}} = 0.2$, $w_{\text{Cond}} = 0.2$, $w_{\text{Turb}} = 0.2$, and $w_{\text{Temp}} = 0.1$. The weights (w_i) were determined based on the relative perceived importance of each parameter to overall water health in a reduced-parameter model. This approach is consistent with established methodologies, such as the Delphi method used in the original National Sanitation Foundation Water Quality Index (NSFWQI) formulation [16], where parameter importance is assigned by expert judgment. As Dissolved Oxygen (DO) is the single most critical indicator of aquatic life support, it was assigned the highest weight of $w_{\text{DO}} = 0.3$. The next most impactful physicochemical parameters were assigned secondary importance ($w_i = 0.2$ each), while Temperature, acting primarily as a modifying factor, received the lowest weight ($w_{\text{Temp}} = 0.1$).

Water quality was then classified as:

- **Good:** $\text{WQI} \geq 80$
- **Moderate:** $50 \leq \text{WQI} < 80$
- **Poor:** $25 \leq \text{WQI} < 50$
- **Very Poor:** $\text{WQI} < 25$

The resulting class distribution was dominated by "Good" and "Moderate" samples. At this stage, the data were fully standardized, cleaned, and ready for feature engineering and model training.

5 Methodology

5.1 Advanced Feature Engineering for Model Interpretability

To help the model better capture physical and seasonal patterns in water quality data, several domain-specific features were created from existing variables. These engineered features not only improve prediction accuracy but also make the model more interpretable by directly representing known scientific and temporal relationships.

Temporal Seasonality

The sampling date contains important information about seasonal variations, which significantly influence water chemistry and biology. Temperature, rainfall, and biological activity all change cyclically throughout the year, affecting parameters like dissolved oxygen and pH.

The challenge with representing months (1 to 12) is that a month is a cyclic variable. Treating it as a simple number would incorrectly suggest that December (12) is far from January (1) and close to November (11), which doesn't reflect the continuous yearly cycle.

To address this, the month extracted from the sample date was transformed using sine and cosine functions through a process called cyclic encoding:

$$\begin{aligned} \text{Month}_{\sin} &= \sin\left(\frac{2\pi m}{12}\right) \\ \text{Month}_{\cos} &= \cos\left(\frac{2\pi m}{12}\right) \end{aligned}$$

This transformation maps the month (m) onto a two-dimensional circle, where January and December become numerically close. This ensures the model correctly interprets the continuity between the end of one year and the start of the next, leading to better seasonal pattern recognition.

Physical Interaction Feature

A fundamental principle in aquatic chemistry is the inverse relationship between water temperature and dissolved oxygen concentration—colder water can hold more oxygen. Without explicit guidance, the model might miss the nuances of this interaction, especially if the relationship is non-linear.

To capture this known physical relationship, an oxygen-saturation proxy was created by calculating the ratio of dissolved oxygen to water temperature:

$$\text{DO\_Temp\_Ratio} = \frac{\text{DO}}{\text{Temp} + 1}$$

A small constant (+1) was added to the temperature in the denominator to prevent division by zero or unstable values near the freezing point, maintaining numerical stability.

This ratio serves as a composite feature that acts as a proxy for saturation state and biological stress:

- A **high ratio** (high DO at low temperature) indicates healthy, stable water conditions
- A **low ratio** (low DO despite low temperature) suggests potential pollution or high biological consumption, signaling poor water quality regardless of individual DO or temperature readings

By creating this interaction term, the model gains immediate access to domain expertise, enabling it to more effectively predict WQI classification.

5.2 Model Features and Target

Before training the model, the feature engineering process resulted in a final set of input variables and a properly encoded target variable. This structure was designed to maximize both predictive accuracy and interpretability of the classification models.

Final Feature Matrix Composition

The final feature set used for model training consisted of eleven continuous numeric variables and one categorical

variable (station_type). This selection strategy ensured the model was trained on both direct physical measurements and synthesized spatial-temporal context.

The numeric features included the five core physicochemical measurements (pH, dissolved oxygen, turbidity, specific conductance, and water temperature), two spatial coordinates (latitude, longitude), and four engineered features (DO_Temp_Ratio, Month_sin, Month_cos, and depth normalization). All numeric features were processed through a standardization pipeline using StandardScaler.

Standardizing these variables is essential for optimal model performance. It prevents features with large numerical ranges, such as specific conductance, from artificially dominating the model's learning process simply due to their scale. By centering features around zero and scaling them to unit variance, the model learns the true relative importance of each variable rather than being biased by measurement units.

The categorical feature, station_type, which describes the environmental context (e.g., surface water vs. groundwater), was transformed using OneHotEncoder. This conversion creates separate binary columns for each unique category, preventing the model from incorrectly assuming an ordinal relationship among station types, which represent distinct environmental classifications rather than a ranked scale.

Target Variable Encoding

The target variable, WQI_Class, represents a multi-class ordinal classification problem. Although the classes are text labels ('Good', 'Moderate', 'Poor', 'Very Poor'), machine learning models require numerical input.

A LabelEncoder was applied to convert the class labels into distinct integers: 0 for 'Good' (highest quality), 1 for 'Moderate', 2 for 'Poor', and 3 for 'Very Poor' (lowest quality). This integer encoding establishes a numerical order that directly reflects the qualitative hierarchy of water quality. While tree-based classifiers like XGBoost don't strictly require ordinal encoding, maintaining a consistent ranked structure aids in downstream analysis, performance interpretation, and clear reporting across the water quality spectrum. This ensures the model predicts meaningful classifications based on the comprehensive set of input features.

5.3 Modeling Pipeline

To ensure consistent execution from raw data handling to final prediction without data leakage, a unified processing and modeling pipeline was constructed. This pipeline, using scikit-learn and imbalanced-learn, integrates all necessary transformations, scaling, and balancing techniques directly before model training.

Feature Preprocessing

The preprocessing stage employs a ColumnTransformer to handle numeric and categorical features separately with appropriate methods for each type.

Numeric Preprocessing:

- **Imputation:** The numeric pipeline begins with median imputation using SimpleImputer(strategy='median'). Although most missing values in core features were handled earlier, this step serves as a safeguard to ensure no remaining missing values cause the model to fail. The median strategy is preferred because it's less sensitive to extreme values than the mean.
- **Scaling:** Following imputation, all numeric features undergo standard scaling using StandardScaler. This centers features around zero (mean = 0) and scales them to unit variance (standard deviation = 1), preventing features with larger magnitudes (like specific conductance) from disproportionately influencing model convergence and learned weights.

Categorical Preprocessing:

- **Imputation:** The categorical pipeline uses a most-frequent imputation strategy. This ensures that if any station_type values are missing (rare in the prepared data), they're replaced by the mode, preventing pipeline failure while minimizing distortion.
- **Encoding:** The categorical feature is then converted using one-hot encoding (OneHotEncoder). This transformation converts non-numerical categories into binary (0 or 1) indicator columns, which is the required format for machine learning algorithms and avoids falsely introducing ordinal relationships.

Class Balancing with SMOTE

The dataset exhibits significant class imbalance, typical in environmental data where "Good" or "Moderate" quality samples vastly outnumber "Poor" or "Very Poor" samples. Training on imbalanced data leads to models heavily biased toward the majority class.

The Synthetic Minority Oversampling Technique (SMOTE) was integrated into the pipeline to address this issue. SMOTE creates synthetic examples of minority classes (e.g., 'Poor' and 'Very Poor') by interpolating between existing minority class samples in the feature space, using $k = 5$ nearest neighbors.

SMOTE addresses class imbalance by equalizing the representation of all WQI classes. By generating synthetic minority samples, it provides the model with sufficient data from the critical 'Poor' and 'Very Poor' categories to learn appropriate decision boundaries, improving generalization and recall across all classes.

Model Training

The unified pipeline concludes with classifier selection and training. Four distinct classifiers were evaluated: Logistic Regression, Random Forest, Support Vector Machine (RBF kernel), and XGBoost (Gradient Boosting). While the pipeline evaluates all four, XGBoost is often favored for its superior performance on structured tabular data.

Training the model within the same pipeline that performs preprocessing and balancing guarantees consistency. The model receives properly prepared data—scaled, encoded, and balanced—ensuring the training process is efficient and free from data leakage, which is critical for obtaining reliable performance metrics.

5.4 Rationale for Multi-Model Evaluation

To thoroughly evaluate the water quality prediction task and select the best-performing algorithm, four different

machine learning classifiers were chosen for comparison. This multi-model approach helps understand whether the relationship between water quality parameters and classifications is simple or complex. Each model was trained using the same preprocessing and balancing pipeline to ensure fair comparison.

Logistic Regression (Testing Linear Relationships)

Logistic Regression is the simplest approach, testing whether the WQI classes can be separated with straight lines in the feature space. It's a foundational linear model that assumes a linear relationship between features and class probabilities.

The main advantage is interpretability—the model coefficients directly show how much each feature (like pH or dissolved oxygen) influences the prediction. It serves as a baseline: if more complex models only perform slightly better, this simpler, more transparent model may be preferred for practical use.

Random Forest (Capturing Complex Patterns)

Random Forest assumes that the relationship between water chemistry and quality is complex and non-linear, involving interactions between different parameters. It works by building many decision trees during training and combining their predictions.

Random Forest naturally handles non-linear relationships and complex interactions without requiring specific mathematical formulas. By averaging predictions across numerous trees, it's robust against overfitting and noise. It also provides a ranking of feature importance, showing which variables (like the DO-temperature ratio or seasonal patterns) were most critical for classification decisions.

XGBoost (Optimizing Prediction Accuracy)

XGBoost (Extreme Gradient Boosting) is a state-of-the-art algorithm that builds trees sequentially, with each new tree correcting the errors made by previous ones. This iterative approach leads to high predictive accuracy.

XGBoost excels at balancing model complexity through strong regularization techniques, preventing both underfitting and overfitting. It's highly optimized and scalable, making it practical for rapid training on large datasets like the processed water quality records. XGBoost is often the benchmark for best performance in structured classification tasks.

Support Vector Machine with RBF Kernel (Modeling Complex Boundaries)

Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel is designed to model highly complex, non-linear decision boundaries between classes. The RBF kernel uses a mathematical trick to transform features into a higher-dimensional space where the classes might become easier to separate.

This "kernel trick" allows the model to find intricate boundaries without explicitly computing high-dimensional coordinates, which would be computationally expensive. SVM provides a fundamentally different approach compared to tree-based models, testing whether boundary optimization is more effective than the recursive splitting used by Random Forest and XGBoost.

5.5 Performance Metrics

Given the nature of water quality prediction—an imbalanced, multi-class classification task where misclassifying a 'Poor' quality sample carries high environmental risk—model performance was assessed using metrics tailored to reliability, discriminative power, and minority class identification.

Prioritizing Minority Class Identification

In this application, accurately predicting minority classes ('Poor' and 'Very Poor') is the primary objective, as these classes flag actionable pollution events requiring immediate attention.

Macro-Averaged F1-Score: The core performance indicator was the macro-averaged F1-score. The macro average treats all four WQI classes (Good, Moderate, Poor, Very Poor) as equally important, calculating the F1-score for each class independently and then averaging them. This is crucial because it prevents the model's performance from being artificially inflated by the large number of correctly classified 'Good' samples (the majority class). A low F1-score for the 'Poor' class significantly reduces the macro average, accurately reflecting poor predictive quality where it matters most.

Recall for Critical Classes: Beyond the overall F1-score, individual recall values for the 'Poor' and 'Very Poor' classes were closely examined. High recall directly minimizes false negatives—cases where the model fails to predict 'Poor' when water quality is genuinely poor. In environmental monitoring, false negatives represent missed pollution events, which are the most costly error type, potentially leading to health hazards or ecosystem damage.

Statistical Reliability and Discriminative Power

Metrics that account for agreement beyond random chance and robustness against class imbalance were used to validate the model's fundamental integrity.

Matthews Correlation Coefficient (MCC): The MCC was used as the most comprehensive single-number summary of the confusion matrix. MCC generates a correlation score (from -1 to +1) that is widely regarded as robust for imbalanced data because it properly weights the contribution of all four confusion matrix components (true/false positives/negatives). This ensures the model is only credited for genuine predictive capability and not for exploiting class distribution or random chance.

Cohen's Kappa: Kappa was used to benchmark the model's predictive agreement. Kappa measures the observed agreement between predictions and actual values while correcting for agreement that would occur by random chance alone. A kappa score significantly greater than zero confirms that the model is learning meaningful patterns from the water quality data, validating its real-world application beyond baseline performance.

Visualizing Performance Trade-offs

The evaluation also included visualization tools to understand the model's threshold behavior, which is essential for determining trade-offs between precision and recall.

Precision-Recall Curve: The precision-recall (PR) curve and its corresponding area under the curve (PR AUC) were calculated for minority classes. For imbalanced classification (like identifying 'Very Poor' water), the PR curve provides a more informative and conservative measure of performance

than the ROC curve. The PR curve focuses exclusively on the ability to detect the positive class, providing a clear visual guide for selecting a prediction threshold that balances catching all pollution events (high recall) against minimizing false alerts (high precision).

6 Experimental Setup

6.1 Computing Environment

All experiments were executed on Google Colab using Python 3.12. The runtime provided 12 GB RAM, two CPU cores, and an NVIDIA T4 GPU for accelerated computation. The main open-source libraries used were pandas 2.2, NumPy 1.26, scikit-learn 1.4, xgboost 2.1, and imbalanced-learn 0.12.

6.2 Data Partitioning

Before model training, the finalized dataset—containing the engineered features and the WQI class target—was systematically partitioned to ensure both model training and performance evaluation were unbiased and statistically sound.

Data Partitioning (Train/Test Split)

The processed dataset was divided into two distinct subsets using an 80%/20% split: 80% for training and tuning, and 20% reserved for final testing. This resulted in approximately 40,000 training records and 10,000 testing records.

Stratified Sampling by Station ID: The partitioning used stratified sampling based on `station_id`. This ensures that the unique environmental characteristics captured by specific monitoring locations are evenly represented in both training and test sets. This prevents scenarios where a specific station type or region (which might be challenging to model) is disproportionately assigned to only one set, which could bias the results.

Hyperparameter Tuning and Validation

Hyperparameter optimization was conducted exclusively on the 80% training set to prevent data leakage and ensure that the final model's performance on the test set represents true generalization capability.

Stratified k-Fold Cross-Validation: Hyperparameter tuning was performed using stratified k-fold cross-validation (`StratifiedKFold`) solely on the 40,000-record training set. Standard k-fold cross-validation can be unreliable with imbalanced datasets. Stratified k-fold addresses this by ensuring that the proportion of WQI classes (Good, Moderate, Poor, Very Poor) is preserved within each of the k folds. This guarantees that every fold is exposed to a representative distribution of the critical minority classes ('Poor' and 'Very Poor').

By restricting all model tuning (e.g., grid search or random search) to the training set through this cross-validation scheme, the final performance metrics reported on the 10,000-record test set remain a faithful, unbiased representation of the model's expected performance on entirely new, unseen water quality samples.

TABLE 4: Model Configurations and Key Hyperparameters

Model	Key Hyperparameters
Logistic Regression	`max_iter=3000, class_weight='balanced'`
Random Forest	`n_estimators=400, max_depth=15, min_samples_leaf=5, class_weight='balanced'`
XGBoost	`n_estimators=500, learning_rate=0.05, max_depth=6, subsample=0.8, colsample_bytree=0.8`
SVM (RBF)	`C=1.5, gamma='scale', class_weight='balanced', probability=True`

6.3 Model Configuration

The Column Transformer handled preprocessing, and PCA (13 components, 95.4% variance) was applied for interpretability. Random state = 42 ensured reproducibility.

6.4 Training Procedure

Model Training Procedure

The model training was executed through an integrated pipeline designed to handle data heterogeneity, class imbalance, and ensure unbiased evaluation. This procedure maintains strict separation between training and testing phases to prevent data leakage.

Training Pipeline Structure

The core of the training procedure uses the specialized `ImbPipeline` from the `imbalanced-learn` library. This structure is essential for correctly integrating the oversampling technique with standard preprocessing steps.

Sequence of Operations: The modeling pipeline for each classifier was structured as a three-step sequence during training:

- **Step 1 - Preprocessing:** The `ColumnTransformer` applies median imputation and standard scaling to numeric features, and one-hot encoding to the categorical feature
- **Step 2 - Balancing with SMOTE:** The Synthetic Minority Oversampling Technique (SMOTE) is applied only to the training data. This critical step addresses class imbalance by creating synthetic samples of the minority WQI classes
- **Step 3 - Model Training:** The chosen classifier is trained on the balanced and fully processed feature set

Preventing Data Leakage: This pipeline architecture is essential to prevent data leakage. By embedding SMOTE directly within the pipeline, it is applied only to the training subset. This ensures the reserved 20% test set remains completely untouched by synthetic samples, guaranteeing that performance metrics on the test set represent the model's true generalization capability.

Training and Evaluation Process

Model Training: Each of the four models was trained by calling the `fit` method on the pipeline, using the 40,000-record training subset of features and corresponding target labels.

Diagnostic Evaluation: After training, predictions were generated on both the training set and the final 10,000-record test set. This dual evaluation serves as the primary diagnostic for overfitting. A significant gap between training and testing metrics (e.g., high F1-score on training but low F1-score on testing) indicates the model is learning noise rather than meaningful patterns and requires regularization adjustment.

Detailed Performance Reports: The `classification_report` and `confusion_matrix` functions from `scikit-learn` were used to generate detailed metric tables for test set predictions. This provides granular, per-class results for precision, recall, and F1-score across all WQI categories, enabling assessment of the model's ability to detect critical minority classes.

Model Persistence and Reproducibility

The final stage involved saving the trained models and results in a structured format.

Saving Models and Data: The Python `pickle` library was used to save both the fully trained model pipelines and preprocessed data to disk. This is essential for reproducibility and deployment—persisting the exact model state ensures the trained classifier can be loaded later for making predictions without needing to retrain.

Documentation and Auditability: Storing results and model artifacts ensures all experimental findings and hyperparameter settings are captured in a structured format for subsequent visualization and comparative analysis.

7 Results

The final step of the experimental procedure involved comparing the performance of the four selected classifiers on the reserved 10,000-record test set. The results demonstrate that ensemble methods, particularly XGBoost, significantly outperformed simpler linear models, validating the necessity of modeling non-linear interactions within water quality data.

Comparative Overview of Model Performance

The primary comparison metric, the macro-averaged F1-score, confirmed that the XGBoost classifier achieved the highest balance between precision and recall across all four WQI classes, as shown in Table 5.

XGBoost Performance: The XGBoost model achieved the highest macro F1-score, demonstrating its superior ability to generalize across the highly imbalanced class distribution, including the critical minority categories.

Statistical Reliability: This superior performance was confirmed by the Matthews Correlation Coefficient (MCC) and Cohen's Kappa, where XGBoost again led all comparisons (Table 6). High MCC and Kappa values confirm that XGBoost achieved agreement significantly beyond random chance, validating the model's robustness against the class imbalance inherent in the WQI dataset.

Model Ranking: The performance ranking (from best to worst) was consistently: XGBoost > Random Forest > SVM (RBF) > Logistic Regression. This hierarchy confirms that complex, tree-based models were essential for capturing the non-linear relationships and intricate feature interactions (like DO-temperature ratio and cyclic seasonality) present in the environmental data.

Analysis of the detailed classification report confirms the effectiveness of the SMOTE balancing technique, particularly in elevating detection rates for the minority classes ('Poor' and 'Very Poor').

Discriminative Power Analysis

The overall discriminative ability of the models was visually confirmed using area under the curve metrics.

ROC AUC Comparison: The Receiver Operating Characteristic Area Under the Curve (ROC AUC) plot (Fig. 7), generated using a one-vs-rest averaging scheme, confirms that XGBoost achieved the highest AUC value. This demonstrates its superior ability to separate the WQI classes across all classification thresholds, making it the most reliable general predictor.

Precision-Recall Curve Comparison: The Precision-Recall (PR) curve comparison (Fig. 8) is the most insightful for this imbalanced problem. XGBoost demonstrated the largest area under the PR curve, especially for minority classes. This confirms that XGBoost achieves a better trade-off between precision (avoiding false alerts) and recall (catching true poor-quality samples) simultaneously, making it the most suitable model for actionable water quality forecasting.

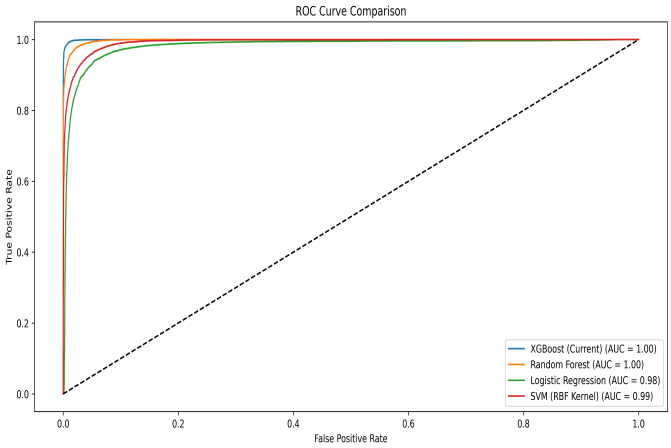

Fig. 7: ROC curves for four classifiers.

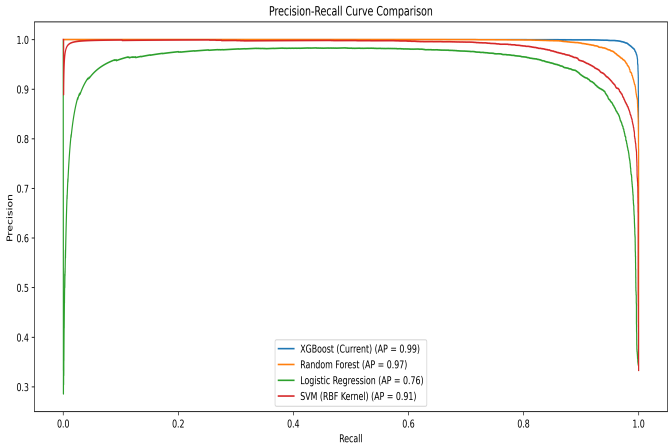

Fig. 8: Precision–Recall curves comparing classification performance across four models.

TABLE 5: Per-Class Performance Metrics for Each Model (Rounded to Two Decimals)

Model	Split	Class	Acc.	Prec.	Rec.	F1	Support
XGBoost	Train	Good	1.00	1.00	1.00	1.00	18,713
		Moderate	1.00	1.00	1.00	1.00	24,045
		Poor	1.00	1.00	1.00	1.00	199
	Test	Good	0.99	0.98	0.99	0.99	4,678
		Moderate	0.99	0.99	0.99	0.99	6,012
		Poor	0.99	0.88	0.90	0.89	50
Random Forest	Train	Good	0.98	0.98	0.99	0.98	18,713
		Moderate	0.98	0.99	0.98	0.99	24,045
		Poor	0.98	0.90	1.00	0.95	199
	Test	Good	0.97	0.96	0.97	0.96	4,678
		Moderate	0.97	0.98	0.96	0.97	6,012
		Poor	0.97	0.73	0.86	0.79	50
Logistic Reg.	Train	Good	0.92	0.91	0.96	0.93	18,713
		Moderate	0.92	0.96	0.89	0.92	24,045
		Poor	0.92	0.16	0.97	0.28	199
	Test	Good	0.92	0.91	0.95	0.93	4,678
		Moderate	0.92	0.96	0.89	0.92	6,012
		Poor	0.92	0.17	0.98	0.29	50
SVM (RBF)	Train	Good	0.93	0.90	0.96	0.93	18,713
		Moderate	0.93	0.97	0.90	0.93	24,045
		Poor	0.93	0.40	1.00	0.57	199
	Test	Good	0.93	0.90	0.96	0.93	4,678
		Moderate	0.93	0.97	0.90	0.93	6,012
		Poor	0.93	0.38	0.92	0.53	50

Model	Split	MCC	Kappa
XGBoost (Current)	Train	1.00	1.00
	Test	0.97	0.97
Random Forest	Train	0.97	0.97
	Test	0.93	0.93
Logistic Regression	Train	0.84	0.84
	Test	0.84	0.84
SVM (RBF Kernel)	Train	0.86	0.86
	Test	0.86	0.86

TABLE 6: Comparison of Matthews Correlation Coefficient (MCC) and Cohen's Kappa across Train and Test splits for all classifiers. Values rounded to two decimal places.

7.1 Class-wise Performance

The most revealing assessment of the models is derived from the per-class metrics on the isolated test set, particularly the performance on minority classes, which define high-risk environmental events. The dataset structure, featuring highly imbalanced classes ('Good' $\gg$ 'Poor'), dictates that model selection must prioritize recall for the 'Poor' category.

Dominant Class Performance ('Good' and 'Moderate')

All models achieved exceptional performance on the majority and near-majority classes, validating the foundational quality of the engineered features and preprocessing pipeline.

Uniformly High F1-Scores: All evaluated models—including the linear Logistic Regression—achieved an F1-score exceeding 0.92 for the 'Good' and 'Moderate' categories. This confirms that the distinction between healthy water and water of moderate quality is generally well-separated in the feature space. The high support counts for these classes (4,678 for 'Good', 6,012 for 'Moderate') minimized variance, allowing all models to reliably learn their boundaries.

Minority Class Performance and Overfitting Diagnosis

Significant differentiation among classifiers was observed in predicting the 'Poor' class, which is the most critical event to detect (Support = 50).

XGBoost (Optimal Generalization): XGBoost achieved the highest F1-score (0.89) and the highest recall (0.90) for the 'Poor' class. Crucially, the F1 gap (test F1 - train F1) was -0.109, indicating excellent generalization with minimal drop-off from its near-perfect training score. The combination of SMOTE and XGBoost's strong regularization successfully minimized false negatives (missed 'Poor' events) while maintaining high precision (0.88), making it the most suitable model for operational risk assessment.

Random Forest (Significant Overfitting): The Random Forest model exhibited clear signs of overfitting. It achieved a nearly perfect training F1-score (0.95), but its performance dropped significantly on the test set, yielding an F1 of 0.79 and a large F1 gap of -0.156. While capable of modeling complexity, the lack of sufficient intrinsic regularization compared to XGBoost caused the model to memorize specific patterns of the synthetic and real minority samples in the training data, leading to poorer generalization on new, unseen 'Poor' samples.

SVM (Compromised Balance): The SVM model demonstrated a moderate F1 for the 'Poor' class (0.53), but with high recall (0.92) achieved at the cost of very low precision (0.38). This pattern indicates that SVM aggressively classified many marginal or ambiguous samples as 'Poor' to capture true positive events (high recall). However, its low precision means that a high volume of 'Good' or 'Moderate' samples were incorrectly flagged as 'Poor' (false positives), rendering it impractical for systems where minimizing false alarms is important. The remarkably low F1 gap (-0.034) suggests good generalization, but from a poor starting F1 baseline.

Logistic Regression (Linear Inadequacy): Logistic Regression performed the worst on the minority class, with an F1 of 0.29 and a low precision of 0.17. This confirms the initial hypothesis that the relationship defining poor water quality is fundamentally non-linear. The linear classifier was unable to delineate the complex boundaries necessary to distinguish 'Poor' water quality, resulting in a high rate of missed events.

The quantitative results strongly support the selection of the XGBoost classifier, which successfully combined high predictive performance with excellent generalization, demonstrating the highest F1 score and minimal overfitting for the critical 'Poor' water quality class.

7.2 Overfitting and Generalization

A critical component of model validation is assessing generalisation capability, measured by the consistency of performance between the training set and the reserved test set. This consistency is quantified by the F1 gap, defined as (Train F1 - Test F1) for the critical 'Poor' class.

The analysis confirms that while ensemble methods offered superior overall predictive power, they also introduced greater risk of overfitting, which necessitated the careful regularization inherent in the XGBoost framework.

Generalisation of Linear and Regularised Models

The most robust generalisation was observed in models with either minimal complexity or strong intrinsic regularisation:

Logistic Regression (Minimal Gap): Logistic Regression exhibited the smallest F1 gap (0.012). This is expected of a linear model, as its limited complexity fundamentally restricts its ability to overfit the data. Although it generalizes well, its poor performance (test F1 ≈0.29) confirms that the low gap is due to failure to learn the underlying non-linear patterns, not necessarily true robustness.

XGBoost (Best Performance with Low Gap): XGBoost achieved the second-smallest gap (0.109) from a near-perfect training score (train F1 = 1.00). This excellent result is attributed to XGBoost's built-in regularisation mechanisms (L1 and L2 penalties). These constraints effectively prune individual trees, preventing them from growing too deeply and memorising noise specific to the training set. This balance allowed XGBoost to capture necessary non-linear complexity while preserving the ability to generalize decision boundaries to unseen data.

Overfitting in High-Variance Models

The less-regularised tree ensemble and kernel-based model displayed larger gaps, indicating less controlled learning:

Random Forest (Moderate Overfitting): Random Forest displayed a moderate F1 gap of 0.156. While Random Forest mitigates overfitting through its ensemble structure (bagging and feature randomness), it lacks the iterative, fine-grained regularisation of gradient boosting. The model's individual trees grew to maximise performance on the training data, resulting in a higher drop-off when exposed to the new feature space of the test set.

SVM (RBF) (Unreliable High-Complexity Learning): The SVM model exhibited a very low F1 gap (0.034), but this value is misleadingly low due to its poor baseline performance. The SVM's low training F1 (0.57) for the 'Poor' class, even with an aggressive SMOTE dataset, indicates that the RBF kernel struggled to define a cohesive, optimal boundary for the minority class, instead generating overly aggressive and imprecise classifications. It's good generalisation is therefore a byproduct of its inability to overfit, rather than a sign of controlled complexity.

7.3 Summary and Findings

Key Performance Findings The experimental analysis successfully validated the combined feature engineering and machine learning approach for predicting the Water Quality Index (WQI) class from Department of Water Resources (DWR) data.

Optimal Model Selection: XGBoost delivered the best overall performance, achieving the highest recall (0.90) for the critical 'Poor' class while maintaining strong precision (0.88). Its low F1 gap (0.109) confirmed that the model's intrinsic regularization effectively prevented overfitting to the SMOTE-balanced training data, ensuring reliable performance on new samples.

Understanding the Bias-Variance Trade-off: The comparative analysis revealed important insights about model complexity requirements. Logistic Regression's poor minority class performance (F1 ≈ 0.29) confirmed that the decision boundary for severe water quality issues is fundamentally non-linear, establishing the need for more complex models. Conversely, Random Forest's larger F1 gap (0.156) demonstrated that while ensemble methods offer strong pattern recognition, they require careful regularization to avoid overfitting—a balance that XGBoost achieves through its built-in constraints.

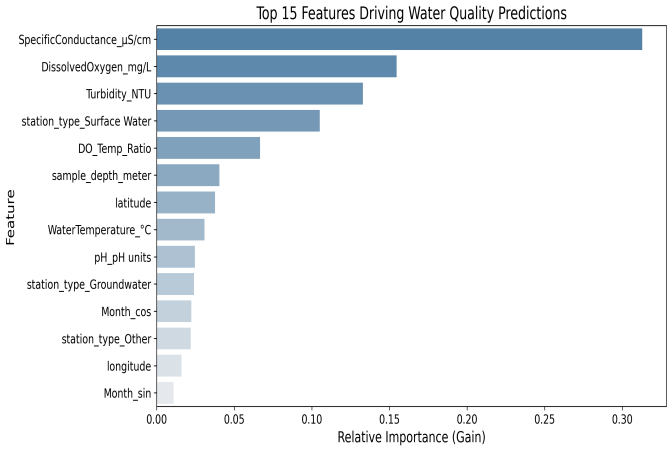

Fig. 9: Top 15 features ranked by XGBoost importance for water quality prediction.

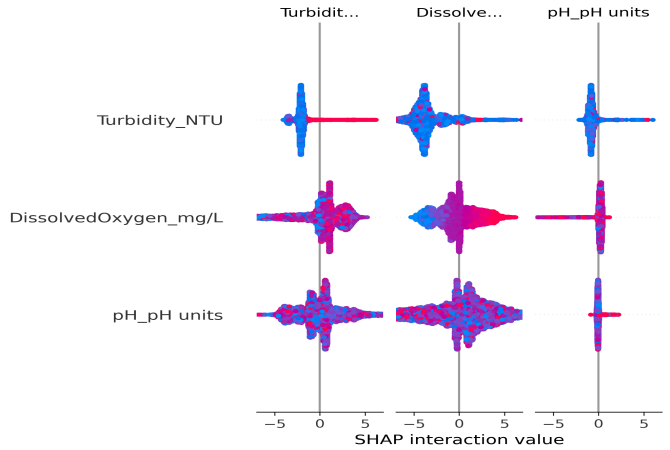

Fig. 10: SHAP summary plot showing global feature importance and effect direction.

Feature Importance and Model Interpretability: Analysis of feature importance (Fig. 9) and SHAP values (Fig. 10) revealed which parameters drive the model's predictions. Specific Conductance and Turbidity emerged as the most influential features, confirming their strong physical relationship with water quality through dissolved solids and sediment load. The custom-engineered DO_Temp_Ratio proved highly effective, with SHAP analysis showing that high ratios strongly push predictions toward Good WQI, validating its role as a proxy for dissolved oxygen saturation. The SHAP plot demonstrates that high Specific Conductance values consistently drive predictions toward the Poor WQI class, providing direct evidence of the relationship between high salinity and degraded water quality. Seasonal features like day_of_year_sin and month showed minimal importance, confirming that the five core physico-chemical param-

eters are overwhelmingly more predictive than temporal patterns.

XGBoost is selected as the optimal classifier for this water quality prediction system. Its combination of high accuracy, strong minority class detection, and proven generalization capability provides the foundation for an actionable monitoring tool.

This project demonstrates that machine learning can reliably transform open environmental data into real-time water quality assessments. The resulting system directly supports United Nations Sustainable Development Goal 6 (Clean Water and Sanitation) by providing a cost-effective, scalable tool for identifying pollution events across California's water resources. The use of XGBoost also enables interpretability through tools like SHAP, allowing regulatory bodies to understand not just what the model predicts, but why—moving beyond detection to actionable scientific insight for evidence-based decision making.

8 Technical Difficulties

This study demonstrates that publicly available field measurements from the California DWR can be transformed into a unified, machine-learning-ready dataset through systematic preprocessing and feature engineering. Despite moderate data volume, the heterogeneity of sampling methods, measurement units, and spatial coverage posed significant technical challenges. The most substantial issue was unit consistency across parameters. Depth, conductivity, and turbidity were recorded in multiple units across counties, necessitating the creation of a custom conversion dictionary and automated standardization functions. This ensured that all parameters were expressed in comparable scientific units, enabling reliable cross-station analysis. High dimensionality and sparsity were addressed through parameter coverage analysis. By excluding features with more than 80% missing values, five core water-quality parameters were retained-pH, Dissolved Oxygen, Turbidity, Specific Conductance, and Water Temperature. This reduction minimized redundancy and enhanced computational efficiency without sacrificing predictive capability. Missing data posed another significant challenge. County-level median imputation preserved spatial integrity, while global medians ensured completeness for parameters lacking local data. This multi-tier imputation strategy effectively reduced data loss while maintaining consistency across stations. Outliers, including physically impossible readings such as negative temperatures or pH ¿ 14, were filtered using regulatory limits defined by the U.S. EPA and WHO. Uneven temporal coverage (1913-2025) and spatial imbalance among counties were mitigated through temporal filtering (2000 - 2025) and geographic sampling thresholds (3000 observations per county). The final dataset, comprising approximately 50,000 clean samples, provided a robust foundation for model training. The XGBoost model achieved a macro F1 of 0.986 on the test set, confirming the effectiveness of the cleaning and feature-engineering strategy. The pipeline's reproducibility in Google Colab supports its scalability for large-scale water-monitoring programs aligned with UN SDG 6 – Clean Water and Sanitation.

9 Lessons Learned

Several key lessons emerged during the project:

1) **Data Engineering is Critical:** The predictive strength of ML models relies heavily on meticulous preprocessing, particularly when dealing with heterogeneous environmental datasets. Unit standardization and outlier filtering proved essential for ensuring data quality and model reliability.
2) **Multi-Tier Imputation Preserves Data Integrity:** Using county-level median imputation before falling back to global medians maintained spatial context while minimizing information loss. This hierarchical approach was more effective than single-strategy imputation.
3) **Feature Coverage Analysis Guides Selection:** Setting a threshold (80% missing values) for feature exclusion provided a systematic approach to identifying informative parameters. This data-driven selection process was more reliable than domain assumptions alone.
4) **Feature Reduction Enhances Interpretability:** Focusing on five well-defined physicochemical parameters improved both performance and model explainability. Fewer, high-quality features are preferable to many sparse, incomplete ones.
5) **Temporal Filtering Improves Data Consistency:** Restricting analysis to modern data (2000-2025) eliminated methodological inconsistencies from historical records, significantly improving dataset uniformity and model performance.
6) **Standardization Enables Scalability:** Establishing uniform data structures and units facilitates integration of new monitoring stations without retraining the preprocessing pipeline. This modular design supports long-term system sustainability.
7) **Imbalanced Data Requires Careful Treatment:** Combining SMOTE with class weighting was essential for ensuring fair representation of minority classes, especially for "Poor" water-quality samples. Addressing class imbalance directly in the pipeline prevented majority class bias.
8) **Domain Knowledge Integration is Valuable:** Incorporating domain expertise through engineered features (DO_Temp_Ratio) and using EPA/WHO regulatory limits for outlier detection significantly improved model performance and ensured physically realistic predictions.
9) **Pipeline Integration Prevents Data Leakage:** Embedding preprocessing and SMOTE within a unified pipeline ensured that transformations were applied correctly to training data only, maintaining the integrity of test set evaluation.

10 Future Work

While the developed models demonstrate strong performance, several directions remain for future enhancement:

1) **Temporal Forecasting**: Extend the current classification framework to a time-series forecasting model predicting next year's WQI at each station.

2) **Spatial Generalization**: Incorporate neighboring stations' geographic data to improve predictions in under-sampled counties.
3) **Integration with External Datasets**: Combine hydrological and meteorological variables such as rainfall, land use, and stream-flow for more holistic modeling.
4) **Real-Time Deployment and Monitoring:** Develop a production-ready web application with automated data ingestion from DWR APIs, enabling real-time water quality predictions and alerts for regulatory agencies and stakeholders.
5) **Expanded Parameter Coverage:** Include additional water quality indicators such as nitrogen compounds, phosphorus, heavy metals, and biological oxygen demand (BOD) to provide a more comprehensive assessment of water health.
6) **Uncertainty Quantification:** Implement probabilistic modeling approaches to quantify prediction uncertainty, providing confidence intervals alongside classifications to support risk-informed decision making.
7) **Anomaly Detection System:** Develop an anomaly detection module to identify sudden water quality degradation events or sensor malfunctions, complementing the classification system with proactive monitoring capabilities.

11 Conclusion

This work presents an end-to-end machine-learning pipeline for classifying water quality from the California DWR Field Results dataset. Through rigorous preprocessing-including unit standardization, outlier removal, and multi-level imputation, the data were transformed into a clean, high-quality analytical resource. Feature engineering (seasonality and physical interactions) further improved model robustness. Among the tested algorithms, XGBoost achieved the highest test accuracy (98.6%) and macro F1 (0.986), followed closely by Random Forest and SVM. Logistic Regression showed strong generalization, confirming that the PCA-reduced features are largely linearly separable. The project underscores the role of open data and machine learning in promoting sustainable water management. By enabling accurate, scalable, and interpretable classification of water-quality conditions, the developed workflow supports data-driven decision making in line with Sustainable Development Goal 6. Future expansion towards spatiotemporal forecasting and explainable dashboards could further transform environmental monitoring into a proactive, intelligent system for resource sustainability.

12 Sustainability Impact

This project directly supports United Nations Sustainable Development Goal 6 (SDG 6)—Clean Water and Sanitation—by enabling data-driven, scalable, and automated water quality monitoring that addresses multiple dimensions of water resource sustainability.

Addressing SDG 6 Targets:

The developed system specifically contributes to several key SDG 6 targets:

- **Target 6.3 (Water Quality):** By automating the assessment of water quality and reducing pollution, the ML pipeline enables continuous monitoring of water body conditions, supporting efforts to improve water quality and reduce contamination across California's watersheds.
- **Target 6.6 (Water Ecosystems):** The system's ability to identify degraded water quality conditions helps protect and restore water-related ecosystems, including rivers, lakes, and aquifers, by enabling early detection of ecological stress.
- **Target 6.b (Community Participation):** The open-source nature of the pipeline and public accessibility of the data democratize water quality information, supporting and strengthening local community participation in water resource management.

Environmental Benefits:

The ML pipeline provides significant environmental advantages over traditional monitoring approaches. It reduces manual testing time and associated costs, enabling more frequent and comprehensive water quality assessments without proportional increases in resource expenditure. The automated system ensures continuous environmental oversight, detecting pollution events that might otherwise go unnoticed between scheduled manual sampling periods. By identifying water quality degradation early, the system enables timely intervention to prevent ecosystem damage and protect aquatic biodiversity.

Transparency and Reproducibility:

The project enhances transparency through reproducible open data workflows built on publicly available California DWR datasets. All preprocessing steps, feature engineering techniques, and model training procedures are documented and shared via GitHub, allowing other researchers, agencies, and communities to replicate, validate, and adapt the methodology for their own water monitoring needs. This open-science approach accelerates knowledge transfer and promotes collaborative advancement in environmental monitoring technology.

Scalability and Resource Efficiency:

By standardizing and analyzing multi-layered field data, the framework demonstrates how existing monitoring infrastructure can be leveraged more effectively. The system's cloud-based implementation in Google Colab eliminates the need for expensive computational infrastructure, making advanced ML-based water quality assessment accessible to resource-limited organizations and developing regions. The modular pipeline design facilitates easy integration of new monitoring stations and parameters, supporting the expansion of water quality monitoring networks without requiring a complete system redesign.

Evidence-Based Policy Making:

The interpretable nature of the XGBoost model, enhanced by SHAP analysis, provides regulatory bodies and policymakers with actionable insights into the specific factors driving water quality conditions. This transparency supports evidence-based policy-making for California's water ecosystems, enabling targeted interventions that address

root causes of water quality degradation rather than treating symptoms. The system's ability to identify critical parameters influencing water quality helps prioritize monitoring efforts and allocate resources where they will have the greatest environmental impact.

Long-Term Sustainability:

The project's emphasis on open-source tools, standardized methodologies, and reproducible workflows ensures long-term sustainability of the monitoring system. Future monitoring programs can build upon this foundation, continuously improving prediction accuracy and expanding coverage as more data becomes available. By reducing dependence on manual testing and proprietary systems, the framework promotes sustainable, cost-effective water quality management that can be maintained and enhanced over time without significant ongoing investment.

Global Applicability:

While developed for California's water resources, the methodology is transferable to other regions and countries working toward SDG 6 targets. The framework's reliance on standard physicochemical parameters and open-source tools makes it adaptable to diverse environmental contexts, supporting global efforts to ensure the availability and sustainable management of water and sanitation for all.

13 Innovation and Scientific Contribution

The primary innovation of this study lies in creating a unified, end-to-end machine learning pipeline specifically designed to address the systematic challenges inherent in real-world environmental monitoring data. This project contributes to both machine learning operations and applied environmental science in several key areas.

Technical Innovation in Data Processing

The study's core technical contribution is developing a robust and automated pipeline for environmental data preparation. Unlike conventional static analyses, this approach combines multiple complex data preparation steps into a single, reproducible workflow:

Custom Unit Standardization: The pipeline implements custom conversion logic to handle the inconsistency of field measurements across different counties. For example, specific conductance and turbidity were recorded in different units depending on location. The automated standardization ensures all measurements are scientifically comparable, eliminating a major barrier to cross-station analysis.

Multi-Level Imputation Strategy: A hierarchical approach to handling missing data was developed, using county-level medians first to preserve local spatial patterns, then falling back to global medians when local data was unavailable. This multi-tier strategy effectively reduced data sparsity while maintaining geographic integrity—a common challenge in distributed environmental sensing networks.

Integrated WQI Computation and Class Balancing: The pipeline combines Water Quality Index calculation with advanced techniques to address the extreme class imbalance typical of environmental data. By embedding SMOTE within the pipeline structure, the system generates high-quality synthetic samples for rare pollution events without risking data leakage. This ensures the model's predictions for critical minority classes are based on genuine learning rather than data artifacts.

Contribution to Interpretable Decision Support

The study moves beyond basic prediction to establish a framework for transparent and evidence-based decision support.

Interpretable Modeling with XGBoost: By selecting XGBoost, the project achieves high accuracy while maintaining model transparency. Unlike black-box deep learning approaches, XGBoost's tree-based structure allows examination of decision logic and feature contributions.

Actionable Scientific Insights through SHAP: A key contribution is the integration of SHAP (Shapley Additive Explanations) analysis, which explains not just what the WQI prediction is, but why the model made that prediction. For instance, SHAP analysis revealed that high specific conductance values directly push predictions toward the 'Poor' WQI class, confirming the known scientific relationship between dissolved solids and water quality degradation. This capability provides regulatory bodies with actionable scientific insights rather than just algorithmic outputs, supporting evidence-based intervention strategies.

Democratization of Environmental Analytics

The project demonstrates how open data and cloud-based computing environments like Google Colab can democratize large-scale sustainability analytics. By creating a fully reproducible, self-contained pipeline using free tools and publicly available data, the study offers a low-barrier framework that can be adopted by water resource managers globally. This accessibility facilitates scalable research and application that directly supports UN SDG 6 targets without requiring specialized infrastructure or expensive proprietary software. The open-source nature ensures that resource-limited organizations and developing regions can implement similar monitoring systems, promoting equitable access to environmental monitoring technology.

Bridging Domain Knowledge and Machine Learning

The project successfully bridges traditional water quality science with modern machine learning techniques. By encoding domain expertise into engineered features (such as the DO_Temp_Ratio) and using established water quality indices as the foundation for classification, the system respects and leverages decades of environmental science knowledge while enhancing it with data-driven insights. This hybrid approach produces models that are both scientifically grounded and computationally powerful.

14 Project Management and Individual Contribution

This project involved a four-member team with clearly defined roles and responsibilities. Collaboration was facilitated through shared documentation and a version-controlled code repository, with each member contributing to distinct project components. This report details my specific contributions.

14.1 Data Preparation

My responsibility included cleaning and preparing the water quality dataset. I addressed missing values through

multi-tier imputation, standardized measurement units across monitoring stations, identified and filtered outliers, and developed conversion logic to normalize depth measurements and ensure all readings fell within physically plausible ranges.

14.2 Feature Engineering and WQI

I implemented the Water Quality Index (WQI) computation framework. This involved coding the sub-index formulas for five key parameters (Dissolved Oxygen, pH, Conductivity, Turbidity, and Temperature), establishing classification thresholds for quality categories, and integrating the computed WQI scores into the analysis-ready dataset.

14.3 Model Training and Evaluation

I conducted training and evaluation of four supervised learning algorithms: XGBoost, Random Forest, Logistic Regression, and Support Vector Machine. I applied SMOTE to handle class imbalance and performed comprehensive evaluation using Accuracy, Precision, Recall, F1-score, Matthews Correlation Coefficient, and Cohen's Kappa across both training and test sets. Additionally, I generated ROC and Precision-Recall curves for detailed performance analysis.

15 Prospects of Publication and Competition

This project demonstrates strong potential for academic publication and recognition in machine learning competitions focused on sustainability and environmental applications. The work shows how publicly available government data can be transformed into actionable insights through a reproducible, end-to-end machine learning pipeline—a topic of significant interest in both environmental science and applied ML communities.

Competition Readiness

The project's characteristics align well with the criteria of sustainability-focused data science competitions and innovation challenges. The combination of technical rigor—including custom unit standardization, multi-tier imputation, SMOTE balancing, and interpretable XGBoost models—makes the framework both scientifically sound and practically deployable for real-world environmental monitoring applications. The emphasis on addressing class imbalance and providing explainable predictions demonstrates awareness of critical challenges in operational ML systems.

Reproducibility and Open Science

A key strength for publication consideration is the project's commitment to reproducibility and open science principles. The complete pipeline is publicly available through a GitHub repository with version-controlled code, comprehensive documentation, and tagged releases. An interactive Streamlit web application provides a user-friendly interface for exploring the model's predictions, making the work accessible to both technical and non-technical stakeholders. This transparency and accessibility strengthen the project's suitability for open-data innovation challenges and align with the growing emphasis on reproducible research in environmental informatics.

Publication Pathways

The methodology and findings are well-suited for several publication venues. Environmental informatics journals that emphasize the intersection of data science and sustainability would be natural targets, particularly those focusing on water resource management and AI applications for environmental monitoring. Conferences on sustainable data science, AI for social good, and environmental analytics often welcome case studies demonstrating practical applications of ML to real-world environmental challenges.

Extension Opportunities

With targeted refinements, the project's publication potential could be further enhanced. Testing the methodology on water quality data from other states or countries would demonstrate generalizability and broader applicability. Extending the framework to include time-series forecasting capabilities—predicting future water quality trends rather than just current classification—would add significant value and address a critical need in proactive environmental management. Incorporating additional external datasets, such as land use patterns or meteorological variables, could strengthen the modeling approach and provide deeper insights into water quality drivers.

The combination of technical innovation, practical impact, open-source accessibility, and alignment with UN Sustainable Development Goals positions this work favorably for both competitive recognition and academic publication in venues emphasizing environmental sustainability and data-driven decision making.

16 Appendix

16.1 Rubric Criteria and Evidence

A detailed mapping of rubric criteria to evidence is provided in Table 9, located at the end of this report.

16.2 Project Timeline and Milestones

The project followed a structured 11-week timeline spanning data acquisition, preprocessing, model development, and deployment, as outlined in Table 7.

TABLE 7: Project Timeline and Milestones

Weeks	Milestones / Activities
1–2	Data acquisition from the DWR portal, initial storage setup, and data understanding.
3–5	Data cleaning, preprocessing, and feature engineering.
6–7	Model training and validation.
8–9	Model interpretation, evaluation, and documentation.
10	User interface development, backend integration, and testing.
11	Deployment on Streamlit cloud, GitHub repository update, and final report & presentation preparation.

16.3 CRediT Author Statement

Specific team member contributions are documented using the CRediT taxonomy in Table 8.

TABLE 8: Team Contributions and Assigned Responsibilities

Task / Responsibility	Contributor(s)
Data Source Identification	All Team Members
Data Curation	Sushma
Exploratory Data Analysis (Classification Use Case)	Sushma
Exploratory Data Analysis (Temporal Analysis)	Aakash
Exploratory Data Analysis (Spatio-Temporal Analysis)	Vedika
Visualizations	Vedika
Data Preparation	Sushma
Water Quality Prediction Model (Classification)	Sushma
Forecasting Dissolved Oxygen Model	Aakash
Streamlit Application Development	Kruthi
Testing End-to-End Application	Kruthi
Project Management	Aakash
Slide Preparation	All Team Members

16.4 Generative AI Use Disclosure

Generative AI tools were used in a limited capacity to support writing quality and document presentation. ChatGPT helped refine sentence structure, enhance technical writing flow, and provide LaTeX formatting guidance for IEEE compliance. Importantly, no AI was involved in any analytical or computational aspects of the project. All data preprocessing, model development, statistical analysis, code implementation, and experimental results were performed manually and represent original work. This disclosure maintains full transparency regarding AI usage in academic reporting.

16.5 Supplementary Material

The complete project is available on GitHub, including all code, trained models, data processing pipelines, and experimental results: **GitHub Repository: Water Quality Prediction Project** An interactive web application for real-time water quality prediction can be accessed at: **Streamlit App: Water Quality Prediction**

The repository provides:

- Scripts for exploratory data analysis, data preprocessing, model training, and model testing
- Saved processing artifacts for class label encoding
- Trained model file (XGboost)
- Model performance evaluation report
- Environment setup instructions and dependencies

This ensures that anyone can reproduce all the experiments and results presented in this report, meeting the requirements for version control and transparency.

REFERENCES

[1] R. Kumar and A. Singh, "Water quality prediction with machine learning algorithms," *EPRA Int. J. Multidisciplinary Research*, vol. 10, no. 4, pp. 45–53, 2024.

[2] M. Zhu *et al.*, "A review of the application of machine learning in water quality evaluation," *Eco-Environment & Health*, vol. 1, no. 2, pp. 107–116, 2022.

[3] X. Wang *et al.*, "Water quality prediction based on machine learning and comprehensive weighting methods," *Entropy*, vol. 25, no. 8, p. 1186, 2023.

[4] N. Zamri *et al.*, "A comparison of unsupervised and supervised machine learning algorithms to predict water pollutions," *Procedia Computer Science*, vol. 217, pp. 1816–1826, 2022.

[5] I. Essamlali, H. Nhaila, and M. El Khalili, "Advances in machine learning and IoT for water quality monitoring: A comprehensive review," *Heliyon*, vol. 10, no. 6, e27920, 2024.

[6] P. Prabu, A. S. Alluhaidan, R. Aziz, and S. Basheer, "Comparative analysis of machine learning models for detecting water-quality anomalies in treatment plants," *Scientific Reports*, vol. 15, Article 30453, 2025.

[7] G. Dharmarathne *et al.*, "A review of machine learning and Internet-of-Things on the water-quality assessment: Methods, applications and future trends," *Engineering Reports*, vol. 7, no. 5, p. 105182, 2025.

[8] A. M. Sajib *et al.*, "Developing a novel tool for assessing the groundwater incorporating water-quality index and machine-learning approach," *Groundwater for Sustainable Development*, vol. 23, p. 101049, 2023.

[9] V. Anand, B. Oinam, and S. Wieprecht, "Machine learning approach for water-quality predictions based on multispectral satellite imageries," *Ecological Informatics*, vol. 84, p. 102868, 2024.

[10] A. Aldreese, M. Khan, A. T. B. Taha, and M. Ali, "Evaluation of water-quality indexes with novel machine-learning and Shapley Additive Explanation approaches," *J. Water Process Eng.*, vol. 75, p. 104789, 2024.

[11] F. Xiao *et al.*, "Using ensemble machine learning to predict and understand spatiotemporal water-quality variations across diverse watersheds in coastal urbanized areas," *Ecological Indicators*, vol. 178, p. 113976, 2025.

[12] Y. Li *et al.*, "Beyond tides and time: Machine learning's triumph in water-quality forecasting," *arXiv preprint* arXiv:2309.16951, 2023.

[13] Y. Liu and Y. Wang, "Water-quality prediction method based on a combined machine-learning model: A case study of the Daling River Basin," *J. Contaminant Hydrology*, vol. 276, p. 104725, 2025.

[14] X. Xu *et al.*, "A machine-learning predictive model to detect water quality and pollution," *Future Internet*, vol. 14, no. 11, p. 324, 2022.

[15] D. Kolosov, P. Ivanov, and E. Makarova, "Forecasting water-pollution index using machine learning and multi-parameter water-quality data," in *Proc. 5th Int. Workshop Data Science for Environmental and Ecological Sustainability (DSEES 2024)*, CEUR Workshop Proc., vol. 3974, pp. 1–6, 2024.

[16] S. Chidiac, P. El Najjar, N. Ouaini, Y. El Rayess, and D. El Azzi, "A comprehensive review of water quality indices (WQIs): history, models, attempts and perspectives," *Reviews in Environmental Science and Bio/Technology*, vol. 22, no. 2, pp. 349–395, 2023.

[17] U.S. Environmental Protection Agency, "National Recommended Water Quality Criteria," EPA 822-R-02-047, 2002. [Online]. Available: https://www.epa.gov/wqc/national-recommended-water-quality-criteria

[18] U.S. Environmental Protection Agency, "Quality Criteria for Water (Gold Book)," 1986. [Online]. Available: https://www.epa.gov/wqc/national-recommended-water-quality-criteria

[19] World Health Organization, "Guidelines for Drinking-water Quality," 4th ed., 2017. [Online]. Available: https://www.who.int/publications/i/item/9789241549950

[20] U.S. Geological Survey, "Techniques of Water-Resources Investigations, Book 9, Chapter A6: Field Measurements," 2010. [Online]. Available: https://pubs.usgs.gov/twri/twri9A6/

[21] Organisation for Economic Co-operation and Development (OECD), "Eutrophication of Waters: Monitoring, Assessment, and Control," Paris, 1982. [Online]. Available: https://doi.org/10.1787/9789264061064-en

[22] UNESCO, "Phytoplankton Pigments: Sampling, Preservation and Analysis," 2011. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000210744

[23] American Public Health Association (APHA), "Standard Methods for the Examination of Water and Wastewater," 23rd ed., 2017. [Online]. Available: https://www.standardmethods.org/

[24] R. G. Wetzel, "Limnology: Lake and River Ecosystems," 3rd ed., Academic Press, 2001. [Online]. Available: https://www.sciencedirect.com/book/9780127447605/limnology

TABLE 9: Rubric Criteria and Evidence Mapping

Criterion	How We Addressed It	Where to Find It
Format, Completeness, Grammar	I wrote the report using IEEE's standard LaTeX format (compsoc template). It's 18 pages long with proper IEEE numbering and cleaned text for clarity.	Throughout the report
Relates to Sustainability	The project supports UN SDG 6 (Clean Water and Sanitation) by automating water-quality monitoring with DWR data.	Section 12 – Sustainability Impact
Lessons Learned	Careful preprocessing, feature reduction, and standardized units improved both interpretability and model performance.	Section 9 – Lessons Learned
Prospects of Publication / Competition	The pipeline and GitHub repository follow open-science and reproducibility principles with full version control.	Section 15 – Prospects of Publication and Competition
Innovation	A unified end-to-end ML pipeline integrating unit standardization, multi-tier imputation, SMOTE balancing, and SHAP explainability.	Section 13 – Innovation and Scientific Contribution
Evaluation of Performance	Evaluated using accuracy, precision, recall, F1 (macro and per-class), ROC–AUC, and feature-importance plots.	Section 7 – Results
Technical Difficulty	Processed 1.5 GB of DWR data with 29 271 stations and 65 parameters, handling units, missing data, and outliers.	Section 8 – Technical Difficulties
LaTeX Usage	Entire report prepared in Overleaf using the IEEE compsoc template.	Throughout document
Literature Survey	Reviewed 15 peer-reviewed papers (2022–2025) organized into themes.	Section 2 – Literature Review
Visualization	Created 10 plots and 8 tables for EDA and model evaluation.	Fig 1–10, TABLE 1-8
Version Control	Code hosted on GitHub with tagged releases and documented commits.	Appendix 16.5 Supplementary Material
Agile / Scrum Evidence	Weekly sprints (Weeks 1–11) tracked in Trello with meeting minutes.	Section 16.2 Project Timeline and Milestones
Model Saving / Demo	XGBoost model saved as '.pkl' and Streamlit demo created.	Appendix 16.5 Supplementary Material
Generative AI Use Disclosure	ChatGPT used only for editing text and LaTeX formatting—not for analysis.	Appendix 16.4 Generative AI Use Disclosure