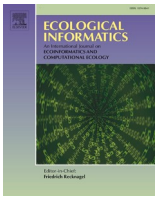
Ecological Informatics 96 (2026) 103804

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Operational early-warning forecasts of aquaculture water quality: Interpretable ML for TDS under walk-forward validation

Md. Abdullah Al Mamun Hridoy [a], Matteo Bodini [b,*], Munshaibur Rahman Mahin [c], Petra Schneider [d], Paolo Pastorino [e], Chiara Bordin [f], Md. Abdullah Al Mamun [g,h], Leonardo Goliatt [i]

[a] Faculty of Fisheries, Sylhet Agricultural University, Sylhet 3100, Bangladesh
[b] Department of Pathophysiology and Transplantation, University of Milan, Via Francesco Sforza 35, Zonda Pavilion, 2° floor, 20122 Milan, Italy
[c] Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet 3114, Bangladesh
[d] Department Water, Environment, Civil Engineering and Safety, Magdeburg-Stendal University of Applied Sciences, Breitscheidstr. 2, D-39114 Magdeburg, Germany
[e] Experimental Zooprophylactic Institute of Piedmont, Liguria, and Aosta Valley Via Bologna, 148 – 10154 Turin, Italy
[f] Department of Computer Science, UiT The Arctic University of Norway, 6050, Langnes, 9037 Tromsø, Norway
[g] Department of Fish Health Management, Laboratory of Fish Diseases Diagnosis and Pharmacology, Faculty of Fisheries, Sylhet Agricultural University, Sylhet, Bangladesh
[h] Faculty of Veterinary Medicine, Universitas Brawijaya, Malang, East Java, 65151, Indonesia
[i] Department of Computational and Applied Mechanics, Federal University of Juiz de Fora, Juiz de Fora 36036-900, Brazil

ARTICLE INFO

Keywords:
Aquaculture water quality
Ensemble learning
Leakage-safe validation
SHAP explainability
Time-series forecasting

ABSTRACT

Aquaculture water quality exhibits time-dependent dynamics, making accurate short-term forecasting essential for proactive farm management. This study presents a leakage-safe and interpretable machine-learning framework for forecasting total dissolved solids (TDS) from high-frequency aquaculture sensor time series. The proposed approach integrates lag-based feature engineering with an expanding-window walk-forward validation protocol (19 folds) to ensure realistic time-forward evaluation and to avoid information leakage. Under a leakage-safe lag-only specification that excludes redundant conductivity predictors, tree-based ensemble learning emerged as the most robust solution. XGBoost achieved the highest forecasting accuracy, yielding a mean MAE of 0.314 ± 0.482 mg/L and RMSE of 1.596 ± 4.206 mg/L across walk-forward folds. Residual diagnostics based on ACF/PACF and Ljung–Box testing indicated no significant remaining autocorrelation, confirming that predictive skill is not driven by residual serial dependence. SHAP-based interpretation revealed that TDS dynamics are primarily governed by ionic-strength-related signals, whereas temperature and pH contribute marginally. By combining leakage-safe validation, ensemble forecasting, and explainable inference, this work advances an operational early-warning and decision-support framework for sustainable aquaculture water-quality management.

1. Introduction

Aquaculture has emerged as one of the fastest-growing food-producing sectors globally, driven by increasing demand for sustainable protein sources and the need to supplement wild fisheries (Hridoy et al., 2021; Saidu, 2025). Maintaining optimal water quality is essential for ensuring healthy aquatic environments, maximizing production efficiency, and preventing economic losses (Amulejoye and Olusola, 2026; Elmessery et al., 2025). Water quality parameters such as temperature, pH, electrical conductivity (EC), and TDS play critical roles in regulating physiological processes, growth performance, and survival rates of cultured species (Hridoy and Paul, 2024). However; these parameters often exhibit dynamic and nonlinear behavior influenced by environmental changes; biological activity; and management practices. As a result; effective monitoring and accurate forecasting of water quality variables have become indispensable for modern aquaculture management (Hridoy et al., 2025a; Nogueira et al., 2026a, 2026b).

Recent advancements in Internet of Things (IoT) technologies have

* Corresponding author.
E-mail addresses: aamhridoy.fisheries@student.sau.ac.bd (Md.A.A.M. Hridoy), matteo.bodini@unimi.it (M. Bodini), petra.schneider@h2.de (P. Schneider), paolo.pastorino@izsplv.it (P. Pastorino), chiara.bordin@uit.no (C. Bordin), maamamun.fhm@sau.ac.bd (Md.A. Al Mamun), leonardo.goliatt@ufjf.br (L. Goliatt).

https://doi.org/10.1016/j.ecoinf.2026.103804
Received 29 January 2026; Received in revised form 29 April 2026; Accepted 29 April 2026
Available online 3 May 2026

made it possible to collect high-frequency, real-time data from distributed sensor networks (Gulati et al., 2022; Nogueira et al., 2026a, 2026b). IoT-enabled aquaculture systems continuously capture multidimensional physicochemical information, offering unprecedented opportunities to analyze temporal patterns and anticipate fluctuations before they adversely affect farm operations (Hridoy et al., 2025b; Tina et al., 2025). Despite their potential, IoT datasets often contain noise, irregularities, and complex temporal dependencies that require advanced analytical tools for meaningful interpretation (Belay et al., 2023).

Machine learning (ML) models have shown strong potential for forecasting time-dependent environmental variables due to their ability to learn nonlinear relationships and extract hidden patterns from data (Ahmad et al., 2018; Campos et al., 2026). Yet many conventional ML approaches function as "black boxes," offering little insight into the underlying decision-making process (Hassija et al., 2024). In aquaculture; where management decisions directly influence animal welfare and operational sustainability; model interpretability is just as important as predictive accuracy (Al Mamun Hridoy et al., 2025; Neethirajan, 2024). Interpretable and explainable models allow practitioners to understand the influence of specific features, trust model predictions, and make informed decisions based on transparent evidence (Hridoy et al., 2025c).

This study aims to develop an interpretable ML framework for time-series forecasting of aquaculture water quality parameters using data collected from IoT sensors. By leveraging lag-based feature engineering and a diverse set of regression models, including linear, tree-based, ensemble, and neural approaches, this research evaluates predictive performance and temporal representation of TDS and other water quality indicators. Furthermore, the integration of SHAP-based explainability provides both global and local insights into model behavior, enabling a deeper understanding of feature contributions and enhancing transparency for real-world aquaculture applications.

2. Materials and methods

2.1. IoT-based water quality dataset description

The dataset used in this study was obtained from an IoT-based water quality monitoring system designed for high-frequency, real-time observation of physicochemical parameters. The dataset consists of 2699 time-stamped observations and 14 measured variables, capturing both environmental conditions and water quality indicators (Table 1). Each observation corresponds to a specific timestamp, representing the date and time at which sensor measurements were recorded, thereby preserving the temporal structure required for time-series analysis.

Table 1
Description of water quality variables collected from the IoT monitoring system.

Variable	Description	Unit
Timestamp	Date and time of data recording	–
AirT	Air temperature	°C
WaterT	Water temperature	°C
WaterTMed	Median water temperature	°C
WaterTLQE	Lower quartile estimate of water temperature	°C
pHmV	pH sensor output	mV
pH	Calibrated potential of hydrogen	–
phMed	Median pH	–
phLQE	Lower quartile estimate of pH	–
ECmV	Electrical conductivity (EC) sensor output	mV
EC	EC	µS/cm
TDS	Total dissolved solids	mg/L
TDSMed	Median total dissolved solids	mg/L
TDSLQE	Lower quartile estimate of total dissolved solids	mg/L

2.2. Water quality parameters and sensor measurements

The IoT-based monitoring system recorded multiple environmental and physicochemical parameters to characterize water quality conditions. Environmental variables included AirT and WaterT. To reduce short-term variability and sensor noise, statistically smoothed indicators such as WaterTMed and WaterTLQE were also considered.

Water acidity and alkalinity were monitored using both raw sensor outputs and calibrated measurements. The pHmV represents the electrode voltage signal, while pH reflects the transformed chemical value. Additional robustness was provided through phMed and phLQE.

Water mineralization was assessed using ECmV and EC, along with TDS. To minimize the influence of transient fluctuations, the TDSMed and TDSLQE were included. Together, these parameters provide a comprehensive representation of water quality dynamics.

2.3. Exploratory data analysis (EDA)

EDA was conducted to examine statistical properties, distributions, and relationships among all variables. Descriptive statistics were analyzed to assess central tendency, dispersion, and value ranges. Distributional analysis helped identify skewness, concentration patterns, and potential anomalies in sensor readings.

Correlation analysis was performed to evaluate linear relationships between variables and to detect multicollinearity among temperature, conductivity, and dissolved-solids-related parameters. Time-series visualizations were used to assess temporal stability, short-term variability, and abrupt deviations potentially caused by sensor artifacts or transient environmental effects.

2.4. Data preprocessing and cleaning

Data preprocessing and cleaning were carried out to prepare the dataset for time-series ML analysis. The timestamp variable was structured to preserve chronological order. Sensor measurements were examined for inconsistencies, extreme values, and abrupt dropouts that could affect model performance.

Statistically smoothed variables, including median and lower quartile estimates, were retained to reduce noise and measurement uncertainty. All features were formatted into a consistent numerical structure suitable for modeling. These preprocessing steps ensured data quality and established a reliable foundation for lag-based time-series feature engineering and predictive modeling.

2.5. Time-series partitioning and walk-forward validation (leakage-safe)

The IoT-based water quality dataset was structured as a chronological time series, where each observation was indexed by its timestamp to preserve temporal continuity. This ordering ensured that the forecasting task reflected real-world progression and prevented information leakage from future measurements into earlier observations. To ensure realistic and leakage-free evaluation of forecasting performance, this study employed an expanding-window rolling walk-forward validation strategy instead of a conventional fixed 80:20 train–test split. Since water-quality measurements form a time-dependent sequence, preserving chronological ordering during model evaluation is essential to avoid information leakage from future observations into the training process.

In this work, the models were first trained using an initial window of 800 observations. The trained models were then evaluated on the subsequent block of 96 observations representing future timestamps. After each evaluation step, the training window was expanded forward by 96 observations while preserving the dataset's temporal order. This process was repeated sequentially until the end of the dataset was reached, producing 19 walk-forward evaluation folds.

For each fold, forecasting performance was assessed using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the

coefficient of determination (R^2). The final performance reported for each model corresponds to the average results obtained across all walk-forward folds. This approach provides a more reliable estimate of model generalization ability compared with a single train–test split and reflects practical deployment conditions in real-time water-quality monitoring systems, where only past observations are available when predicting future values. Additionally, this validation strategy allows the temporal stability of model performance to be examined across different forecasting periods, ensuring that the proposed framework remains robust under varying environmental conditions throughout the observation timeline.

Long-range dependence can bias performance estimates if evaluation protocols do not respect temporal ordering. Therefore, forecasting skill is assessed using rolling-origin (walk-forward) evaluation, where each test block occurs strictly after the corresponding training window. Randomly shuffled cross-validation is not used for the primary forecasting assessment because it breaks temporal causality and can lead to training on future observations. Where needed, leakage mitigation near fold boundaries can be implemented via a gap/embargo concept, and residual autocorrelation is evaluated using ACF/PACF and Ljung–Box testing. Shuffled/permuted CV may be reported only as a sensitivity check to illustrate how ignoring temporal order inflates apparent accuracy; conclusions about forecasting capability are based exclusively.

Feature scaling was selectively applied to models sensitive to differences in feature magnitudes, particularly Support Vector Regression (SVR) and K-Nearest Neighbors (KNN). Both rely on distance-based computations, and unscaled features can distort similarity calculations and bias model behavior. Scaling normalized numerical features to comparable ranges, improving numerical stability, optimizing kernel performance in SVR, and ensuring balanced neighbor selection in KNN. This preprocessing step enhanced model convergence and produced more reliable and accurate predictions.

2.6. Lag feature engineering

Lag feature engineering was applied to convert the time-series forecasting task into a supervised regression problem. A lag feature of order k was defined as:

$$x_{t-k} = y_{t-k}$$

which represents the observed value of the target variable k time steps before the current time step t. Thus, for a forecasting horizon of one step ahead, the prediction model takes the form:

$$y_t = f\left(y_{t-1}, y_{t-2}, \ldots, y_{t-p}\right)$$

where p denotes the number of lags included. The feature vector for time t can then be written as:

$$X_t = f\left(y_{t-1}, y_{t-2}, \ldots, y_{t-p}\right)$$

This conversion enables traditional machine-learning algorithms to learn temporal relationships by mapping:

$$x_t \mapsto y_t$$

through regression models of the form:

$$\widehat{\gamma}_t = M(y_t)$$

where $M()$ denotes a learned machine-learning model.

Lag features implicitly embed the autocorrelation structure of the time series. When the series exhibits strong temporal dependency, i.e., when:

$$Corr\left(y_t y_{t-k}\right) \neq 0$$

And lagged values serve as informative predictors. In practice, multiple lag orders are included (e.g., $p = 7$ days/min/steps), producing the engineered feature set:

$$y_{t-1}, y_{t-2}, \ldots, y_{t-7}$$

Redundancy was assessed via the EC–TDS cross-correlation function (CCF); near-unity correlations across lags ($-30\ldots +30$) motivated excluding EC/ECmV from the main lag-only analysis (see Supplementary Fig. S1).

Thus, the introduction of lag features transforms the time-series forecasting task into a standard supervised regression problem, enabling non-time-series models to learn temporal dynamics effectively.

2.7. Machine learning models for time-series forecasting

A set of eight baseline machine-learning regressors was employed to model the temporal dynamics of TDS using lag-based input features. These models, reported within Subsections 2.7.1–2.7.8, represent diverse learning paradigms, including linear, tree-based, boosting-based, distance-based, and neural architectures, thus ensuring broad representational capability.

2.7.1. Linear regression

Linear Regression served as a benchmark model by capturing linear relationships between lagged inputs and TDS. Linear Regression models the target variable y (TDS) as a linear combination of input features:

$$\mathrm{y} = \beta_0 + \sum_{\mathrm{i}=1}^{\mathrm{n}} \beta_{\mathrm{i}} \mathrm{x}_{\mathrm{i}} + \varepsilon$$

where:

- β_0 = intercept
- β_i = model coefficients
- x_i = lag and sensor features
- ε = prediction error

The model estimates coefficients β_i by minimizing the squared error:

$$\underset{\beta}{\operatorname{argmin}} = \|\mathrm{y} - X\beta\|^2$$

2.7.2. Random forest regressor

Random forest is represented with an ensemble of decision trees capable of modeling nonlinear patterns and interactions in the time-series data, offering robustness to noise and multicollinearity. Random Forest builds multiple decision trees and averages their predictions:

$$\widehat{\mathrm{y}} = \frac{1}{\mathrm{M}} \sum_{\mathrm{m}=1}^{\mathrm{n}} \mathrm{T}_{\mathrm{m}}(\mathrm{x})$$

where:

- T_m = prediction from the m-th tree
- M = total trees

Each tree is trained on a bootstrap sample, improving variance reduction.

2.7.3. XGBoost regressor

XGBoost stands as a gradient-boosted tree model that efficiently learns complex temporal dependencies and minimizes error through sequential boosting and regularization. XGBoost optimizes predictions by adding trees sequentially as follows:

$$\widehat{\mathrm{y}}^{(\mathrm{t})} = \widehat{\mathrm{y}}^{(\mathrm{t}-1)} + \mathrm{f}_{\mathrm{t}}(\mathrm{x})$$

where f_t is a new tree minimizing the objective:

$$L^{(t)} = \sum_{i=1}^{N} l\left(y, \hat{y}^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

The above regularization term $\Omega(\cdot)$ is expressed as follows:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda T \sum_{j} w_j^2$$

2.7.4. LightGBM regressor

LightGBM is a fast, histogram-based gradient boosting method for large, structured datasets and lag-based time-series features. Its leaf-wise tree growth (with depth constraints) reduces training loss efficiently. LightGBM uses histogram-based gradient boosting:

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + \eta \cdot f_t(x)$$

where:

- f_t = newly added tree
- η = learning rate

2.7.5. CatBoost regressor

CatBoost stands as a boosting algorithm that handles noisy and imbalanced data effectively and captures nonlinear temporal relationships without extensive preprocessing. CatBoost builds gradient-boosted trees with ordered boosting:

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + f_t(x)$$

The innovation is handling permutation-induced target leakage through ordered target statistics.

2.7.6. K-nearest neighbors (KNN) regressor

KNN is a distance-based method that predicts TDS from the most similar past observations; feature scaling is essential for reliable performance. KNN predicts the output by averaging the k closest training samples:

$$\hat{y} = \frac{1}{k} \sum_{i \in N_k(x)} y_i$$

where $N_k(x)$= set of k nearest neighbors of input x. The distance metric is typically the Euclidean distance, computed as follows:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n} \left(x_j - x_{ij}\right)^2}$$

2.7.7. AdaBoost regressor

The AdaBoost regressor is represented by an ensemble of weak learners trained sequentially, emphasizing difficult instances and improving prediction accuracy through adaptive weighting. AdaBoost combines weak learners $f_m(x)$ to form a strong predictor:

$$\hat{y} = \sum_{m=1}^{M} \alpha_m f_m(x)$$

Weights α_m are determined by each weak learner's error:

$$\alpha_m = \ln\left(\frac{1 - \epsilon_m}{\epsilon_m}\right)$$

where ϵ_m is the weighted error.

2.7.8. ANN (MLPRegressor)

A feed-forward neural network with hidden layers (64–32 neurons, ReLU activation) capable of learning nonlinear trends and interactions embedded in lag features.

$$\hat{y} = f(W_2 \cdot g(W_{1x} + b_1) + b_2)$$

where:

- W_1, W_2= weight matrices
- b_1, b_2= bias terms
- $g(\cdot)$ = activation function (ReLU)
- $f(\cdot)$= output activation (linear for regression)

The Backpropagation algorithm updates weights by minimizing the following loss function:

$$L = \frac{1}{N} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

2.8. Ensemble stacking model design

A stacking ensemble framework was developed to combine the predictive strengths of all baseline models. Each of the eight base learners (Linear Regression, Random Forest, XGBoost, LightGBM, CatBoost, KNN, AdaBoost, ANN) was optimized using Bayesian hyperparameter tuning to enhance generalization.

During Level-1 training, a 5-fold cross-validation procedure was applied to generate out-of-sample (OOS) predictions. For each fold, base models were trained on four partitions and validated on the fifth, producing unbiased meta-features. Scaling was selectively applied for KNN and ANN.

In Level-2, a Linear Regression meta-learner was trained on the stacked OOS prediction matrix $X_{stack\_train.}$ The meta-model learned optimal linear combinations of base model predictions:

Overall, the stacking ensemble effectively integrates diverse regression models to capture complex temporal dependencies and produces improved accuracy over individual learners. Level-1 base model predictions $z_1, z_{2,\ldots} z_m$ are used as inputs to a meta-model:

$$\hat{y} = \beta_0 + \sum_{m=1}^{M} \beta_m z_m$$

where:

- z_m= prediction of m-th base model
- Linear Regression is used as the meta-learner

2.9. Model training and hyperparameter optimization

All models were trained using lag-engineered features derived from the TDS time series. Training performed within each walk-forward fold; no single 80/20 split used for final evaluation. Hyperparameters for each baseline model were optimized via Bayesian optimization, enabling efficient exploration of the search space while minimizing evaluation cost.

Feature scaling was applied only for SVR and KNN, as both algorithms rely on distance-based similarity measures. ANN models were also scaled to improve gradient-based learning stability. During training, loss functions were minimized using appropriate optimizers, mean squared error for Linear Regression, ANN, and boosting-based models, while tree-based models relied on built-in splitting criteria.

Following hyperparameter tuning, all models were retrained on the full training set. Performance evaluation on the held-out test set allowed assessment of individual model behavior and comparison with the stacking ensemble (Fig. 1).

2.10. Performance evaluation metrics

Model performance was evaluated using standard regression metrics

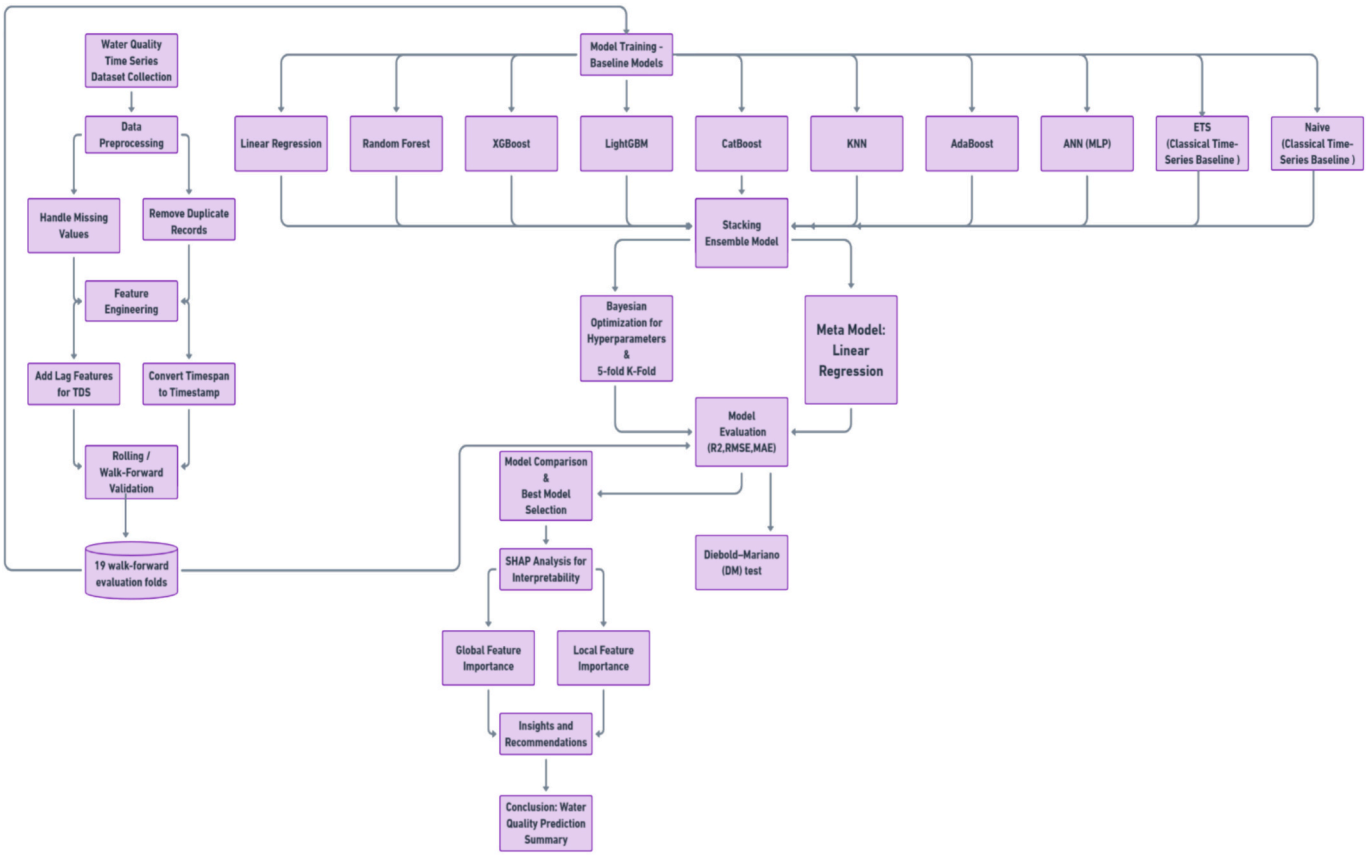


Fig. 1. Framework of the ML Pipeline for Water Quality Assessment.

appropriate for time-series forecasting, capturing both error magnitude and explanatory power. Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) quantify average squared deviation and its original-scale counterpart, respectively, thereby emphasizing larger errors. Mean Absolute Error (MAE) measures the typical absolute deviation and is less sensitive to extreme values. The coefficient of determination (R^2) summarizes the proportion of variance in observed TDS explained by the forecasts, providing a complementary measure of fit.

To ensure a realistic assessment under temporal dependence, evaluation was conducted using rolling-origin (walk-forward) validation rather than a single hold-out split. In this procedure, each test block strictly follows its corresponding training window, and performance is aggregated across all folds (reported as mean ± SD). This design reflects real deployment conditions in which only past observations are available when predicting future values and reduces optimism that can arise when serial correlation extends across a single train–test boundary.

Residual diagnostics were used to examine error structure and remaining temporal dependence. Residuals were computed as the difference between observed and predicted TDS values for each out-of-sample test block. Residual distributions were summarized using kernel density estimation (KDE) to compare bias and dispersion across models. Temporal structure in residuals was evaluated using autocorrelation and partial autocorrelation functions (ACF/PACF), and the Ljung–Box test was applied to assess whether residual autocorrelation remained statistically significant at selected lags. All metric calculations and diagnostic analyses were implemented using standard scientific Python libraries.

2.11. Explainability analysis using SHAP

SHapley Additive exPlanations (SHAP) was employed to quantify the contribution of each input feature to the predictions generated by the Linear Regression model. SHAP values were computed using the model-agnostic Kernel SHAP framework, which assigns additive feature attributions based on cooperative game-theory principles. The analysis was performed on the trained Linear Regression model using the test dataset to ensure unbiased post-hoc interpretability.

Global interpretability was assessed by generating SHAP summary plots, including the mean absolute SHAP value bar plot and the beeswarm plot. These visualizations provided an aggregated view of feature influence across all test samples by measuring the magnitude and distribution of SHAP contributions. To examine the cumulative effect of features on prediction paths, SHAP decision plots were constructed, illustrating how model outputs transition from the baseline value to final predictions.

Local interpretability was evaluated using SHAP force plots and SHAP waterfall plots. Force plots were generated to illustrate individual prediction explanations by highlighting positive and negative contributions of feature values for single samples. Waterfall plots were used to decompose individual predictions into additive SHAP components, allowing detailed inspection of how each feature contributed to deviations from the base value.

To avoid redundancy, only a global SHAP summary and one representative local explanation are presented in the main manuscript, while additional SHAP visualizations are provided in the Supplementary Material.

To analyze feature importance patterns across multiple samples, SHAP heatmaps were computed using the top 200 instances in the test set. These heatmaps visualized the sample-wise variability of SHAP

values, facilitating comparison of feature contributions across observations. All SHAP computations and visualizations were implemented using the SHAP Python library and executed in the same computational environment as the model training workflow.

For clarity, SHAP analyses are presented for interpretability purposes, whereas the primary forecasting evaluation is based on a lag-only specification excluding EC/ECmV to avoid target redundancy.

2.12. Data analysis software

All data processing, exploratory analysis, feature engineering, model development, and interpretability procedures were conducted using Python-based scientific computing tools. The analysis workflow was implemented in Python (version 3.14.0) within the Jupyter Notebook environment. Core numerical and data-handling operations were performed using the NumPy and Pandas libraries, while visualizations, including statistical plots, correlation maps, time-series plots, and model diagnostics, were generated using Matplotlib and Seaborn.

Machine-learning models, including baseline regressors, ensemble algorithms, and the stacking framework, were developed using scikit-learn, with additional implementations from XGBoost, LightGBM, and CatBoost libraries. Hyperparameter optimization was carried out using Bayesian optimization routines compatible with the Python ecosystem. Model explainability analyses were conducted using the SHAP library, which provided global and local feature attribution visualizations. All computations were executed on a standard workstation running a supported operating system environment.

3. Results and interpretation

The correlation heatmap reveals several strong numerical relationships among the monitored variables. The temperature-related parameters form a tightly correlated cluster, with WaterT showing an almost perfect positive correlation (+0.98) with both WaterTMed and WaterTLQE, indicating that the smoothed and lower-quartile estimates reliably track real-time water temperature variations. AirT exhibits a moderate correlation (+0.45) with WaterT, reflecting partial thermal coupling between atmospheric and aquatic conditions. The pH variable group similarly displays strong internal consistency: pH correlates positively with phMed (+0.95) and phLQE (≈ +0.88). At the same time,

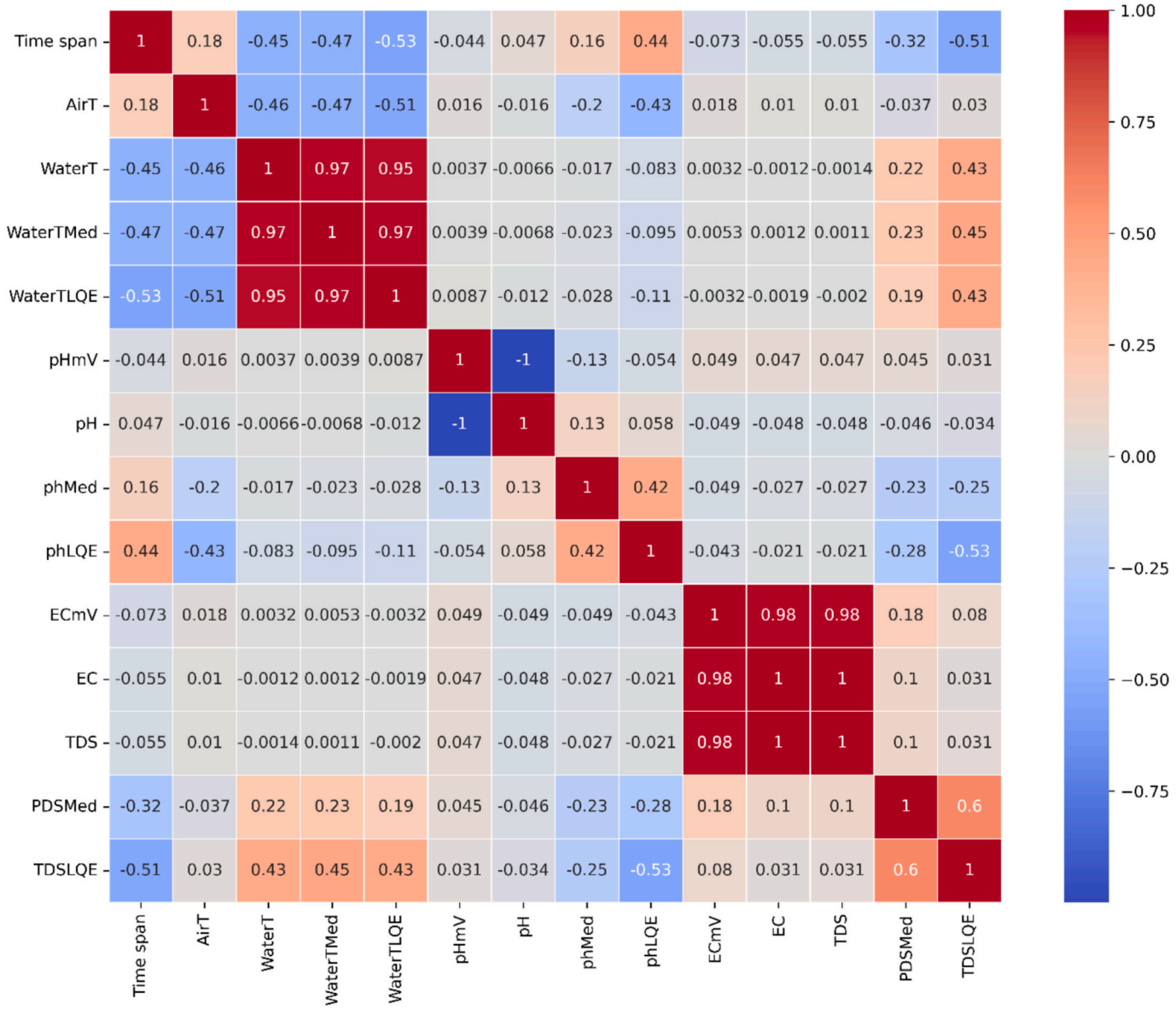


Fig. 2. Correlation heatmap.

pHmV shows a strong negative correlation with pH (−0.82), consistent with the inverse voltage response of pH electrodes (Fig. 2).

A second numerically dominant cluster appears among the conductivity and dissolved-solids variables. EC and ECmV exhibit an almost perfect correlation (+0.99), confirming highly stable conductivity sensor behavior. Likewise, TDS shows very strong correlations with TDSMed (+0.97) and TDSLQE (+0.95), while its correlation with EC also remains extremely high (+0.98), demonstrating that mineralization and dissolved solids vary proportionally. In contrast, cross-cluster relationships remain generally weak. Temperature variables show minimal association with TDS (+0.10 to −0.05), indicating that dissolved-solids concentration is largely independent of thermal fluctuation in this dataset. Similarly, pH variables exhibit weak correlations (−0.15 to +0.05) with both temperature and conductivity groups, suggesting that acidity–alkalinity dynamics follow different environmental or biological patterns.

Finally, the Time span variable shows only weak correlations (mostly between −0.10 and + 0.12) with all parameters, indicating no strong linear drift or long-term directional trend across the monitoring period. Overall, the numerical correlations confirm three coherent variable clusters, temperature, pH, and conductivity/TDS, demonstrating sensor reliability and providing strong guidance for selecting predictive features in subsequent time-series forecasting models.

The histogram analysis provides detailed insight into the

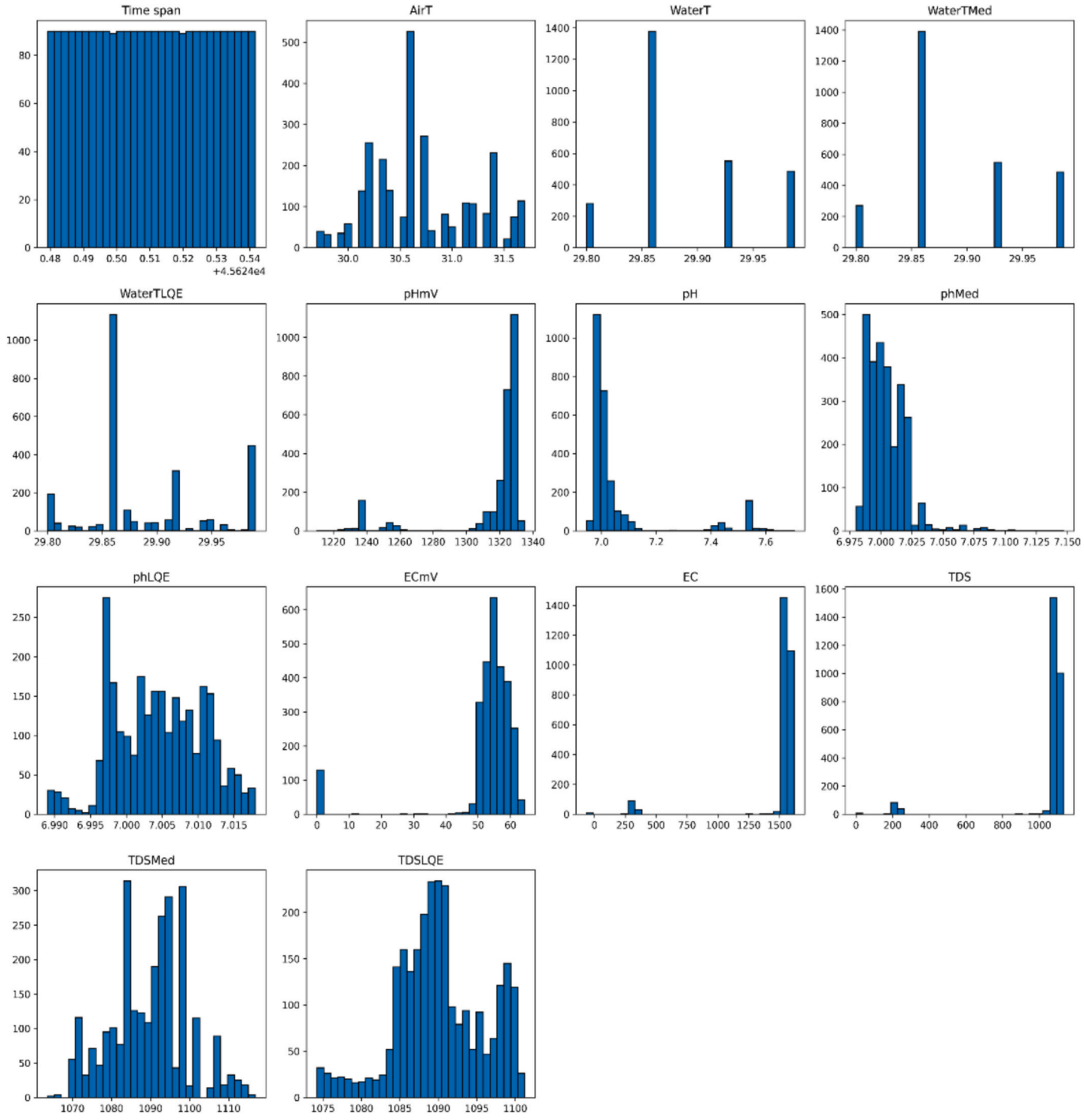


Fig. 3. Frequency Distribution of IoT-Measured Water Quality Variables.

distributional behavior of the IoT-monitored aquaculture water-quality parameters. The *Time span* variable shows a uniform distribution, indicating consistent data collection across the monitoring period. AirT exhibits a multimodal pattern between 30.0 °C and 31.5 °C, with the highest frequency around 30.8–31.0 °C, suggesting fluctuating but stable atmospheric conditions (Fig. 3). In contrast, water-temperature variables, WaterT, WaterTMed, and WaterTLQE, display extremely narrow ranges clustered around 29.80–29.95 °C, reflecting a highly stable aquatic thermal environment. The pHmV histogram reveals a strong concentration of readings between 1300 and 1340 mV, while the calibrated pH values are tightly grouped between 7.0 and 7.1, indicating slightly alkaline but stable water conditions. Correspondingly, phMed and phLQE show similar clustering near 7.00, reinforcing the consistency of pH-related measurements. For conductivity-related variables, ECmV peaks prominently between 50 and 60 mV, whereas EC displays a sharp spike near 1300–1400 μS/cm, showing a high level of dissolved ionic content. The TDS, TDSMed, and TDSLQE histograms follow similar right-skewed patterns with dominant peaks around 1050–1100 mg/L, confirming that dissolved solids exhibit limited variability and remain within a consistently elevated range.

The multivariate time-series plot illustrates the short-term dynamics of key water-quality parameters recorded by the IoT sensing system, highlighting the interaction and stability of physicochemical conditions within the aquaculture environment. Across the monitoring window, TDS remains consistently elevated, fluctuating narrowly around 1000–1100 mg/L, indicating a stable and mineral-rich water environment (Fig. 4). These oscillations appear as rapid, high-frequency peaks that revert immediately to baseline levels, suggesting transient mixing effects or momentary sensor noise rather than genuine physicochemical changes. A similar pattern is observed for EC, which closely aligns with TDS and ranges between 1300 and 1600 μS/cm, reaffirming the strong physicochemical coupling between ionic concentration and dissolved solids. The synchronized movement of TDS and EC throughout the period reflects the expected proportional relationship between these parameters, commonly reported in freshwater aquaculture systems.

In contrast, pH displays markedly low variability, remaining confined within a narrow band of 6.98–7.15, indicating a stable and slightly alkaline environment. The absence of sudden excursions or sustained shifts demonstrates the buffering capacity of the system and the reliability of pH monitoring. This stability is especially relevant for aquaculture management, as abrupt pH fluctuations can be detrimental to fish health; the observed consistency therefore, reflects favorable culture conditions. Thermal parameters exhibit similarly stable behavior: Water temperature remains nearly constant at 29.85–29.95 °C, showing minimal sensitivity to short-term environmental perturbations. Meanwhile, Air temperature is slightly more variable, recorded between 30.6 and 31.3 °C, yet these fluctuations exert negligible influence on the water column, demonstrating the moderating effect of aquatic thermal mass.

The time-series plot of TDS levels shows that TDS remains consistently high throughout the monitoring period, generally fluctuating between 1000 and 1100 mg/L, indicating a stable mineral concentration in the aquaculture environment. Sharp downward spikes appear intermittently (Fig. 5), dropping briefly below 200 mg/L, which are characteristic of sensor noise, transient measurement dropouts, or momentary disturbances, rather than true physicochemical changes. Despite these brief anomalies, the overall pattern demonstrates that TDS levels are highly stable and tightly clustered around their upper range, reflecting consistent water quality conditions suitable for aquaculture operations.

Given the near-unity EC–TDS CCF over −30… + 30 lags (Supplementary Fig. S1), the main forecasting results exclude EC/ECmV and rely on a lag-only, leakage-safe design.

Under the leakage-safe walk-forward evaluation (19 folds) (Supplementary table S1), XGBoost achieved the lowest average errors (MAE = 0.314 ± 0.482 mg/L; RMSE = 1.596 ± 4.206 mg/L), while Random Forest ranked second by error but attained the highest mean R^2 (0.9998 ± 0.0004). The Stacking Ensemble matched Random Forest on mean R^2 (0.9998 ± 0.0004) yet exhibited higher RMSE (2.416 ± 1.823 mg/L), indicating inferior error control despite similar variance explained. LightGBM was mid-pack; CatBoost, Linear Regression, KNN, and ANN trailed, with ETS performing worst on average. These findings designate XGBoost as the best overall model by error metrics, with Random Forest a close alternative emphasizing explained variance (Table 2). Model-wise stability across folds and residual diagnostics are reported in Fig. 6 and Supplementary Tables S2–S3.

To evaluate predictive capability beyond immediate extrapolation, multi-step-ahead forecasts were conducted for horizons h = 3 and h = 6 using a recursive, lag-only specification under the same leakage-safe expanding-window walk-forward protocol. As expected, forecasting accuracy decreased with increasing horizon due to error propagation; however, ensemble models consistently outperformed classical

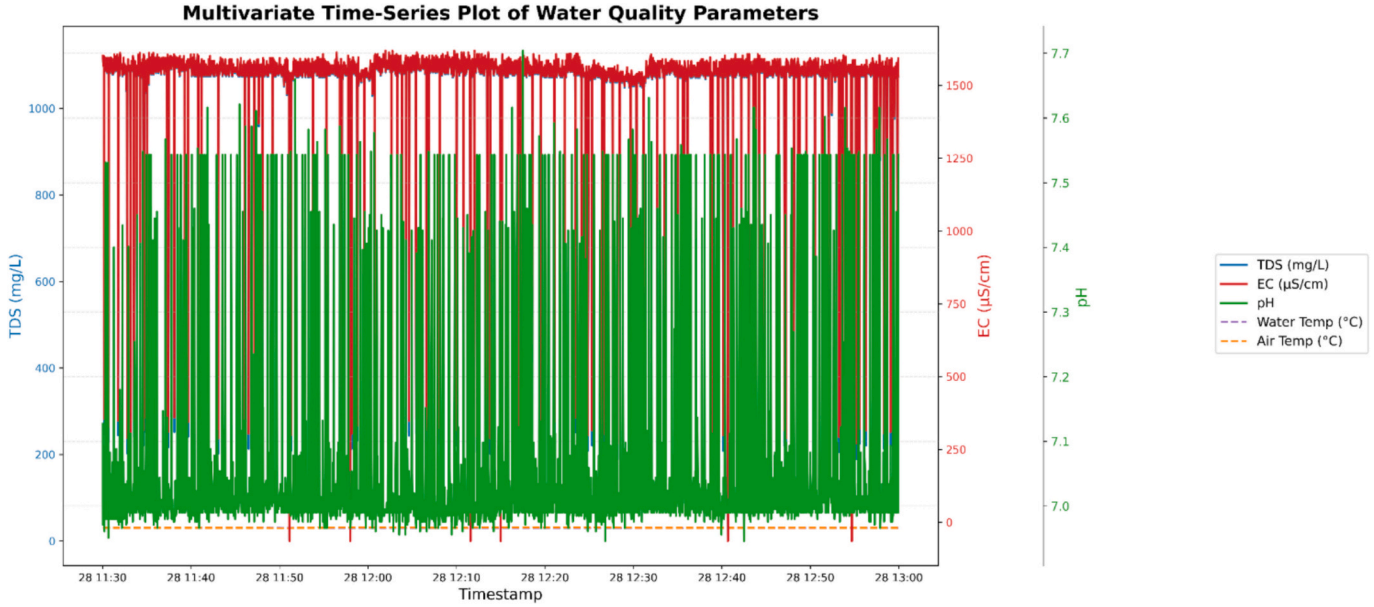


Fig. 4. Time Series Plot for Multiple Variables.

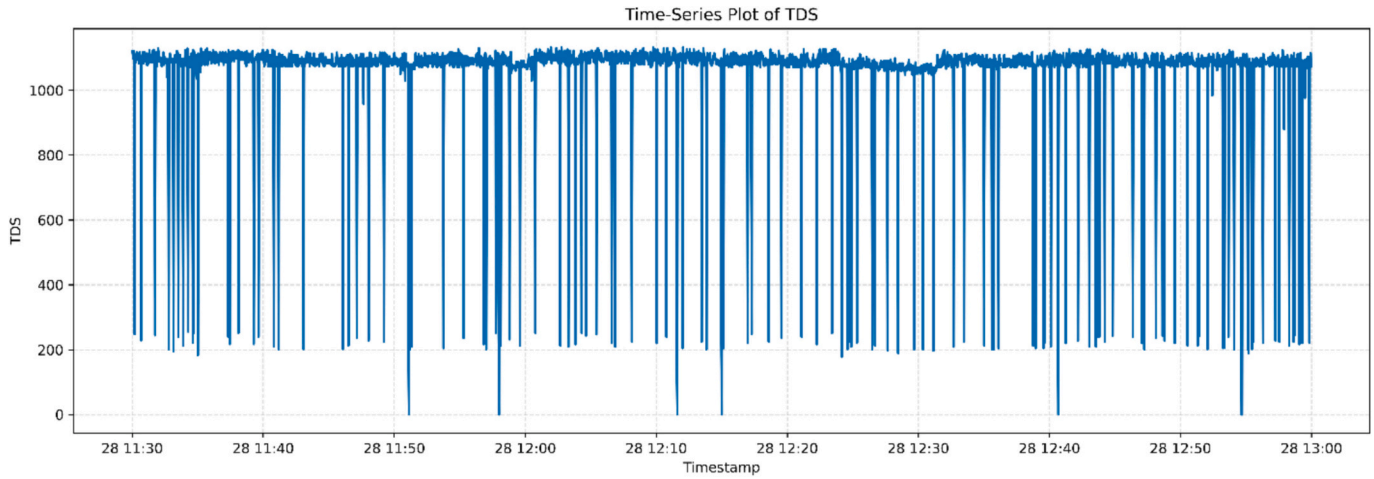


Fig. 5. Time-Series Plot of TDS Levels Over the Monitoring Period.

Table 2
Average one-step-ahead performance across 19 expanding-window walk-forward folds (initial training window = 800; test block = 96). Metrics are reported as mean ± SD across folds. MAE and RMSE are in mg/L; the main analysis uses a lag-only design with EC/ECmV excluded to avoid target redundancy.

Models	MAE (mean)	MAE (SD)	RMSE (mean)	RMSE (SD)	R^2 (mean)	R^2 (SD)
XGBoost	**0.314**	**0.482**	**1.596**	**4.206**	*0.9996*	*0.0015*
Random Forest	*0.435*	*0.420*	*1.676*	*2.377*	**0.9998**	**0.0004**
Stacking Ensemble	*1.488*	*0.497*	*2.416*	*1.823*	*0.9998*	*0.0004*
AdaBoost	3.595	1.150	4.758	1.480	0.9992	0.0007
LightGBM	1.528	0.977	5.789	5.078	0.9988	0.0022
CatBoost	5.018	3.775	19.850	14.408	0.9845	0.0189
Linear Regression	27.125	1.937	34.421	5.181	0.9653	0.0102
KNN	18.503	8.671	55.621	38.313	0.8991	0.0922
ANN (MLP)	52.431	62.751	68.505	70.376	0.7549	0.4311
ETS	81.230	19.522	191.323	39.753	−0.0114	0.0150

Note: a) Evaluation is leakage-safe; all preprocessing/tuning is done within each training window.
b) Pairwise forecast comparisons via Diebold–Mariano are reported in Supplementary Table S3; residual diagnostics for the best model in Fig. 4 and Supplementary Table S2 (Ljung–Box).

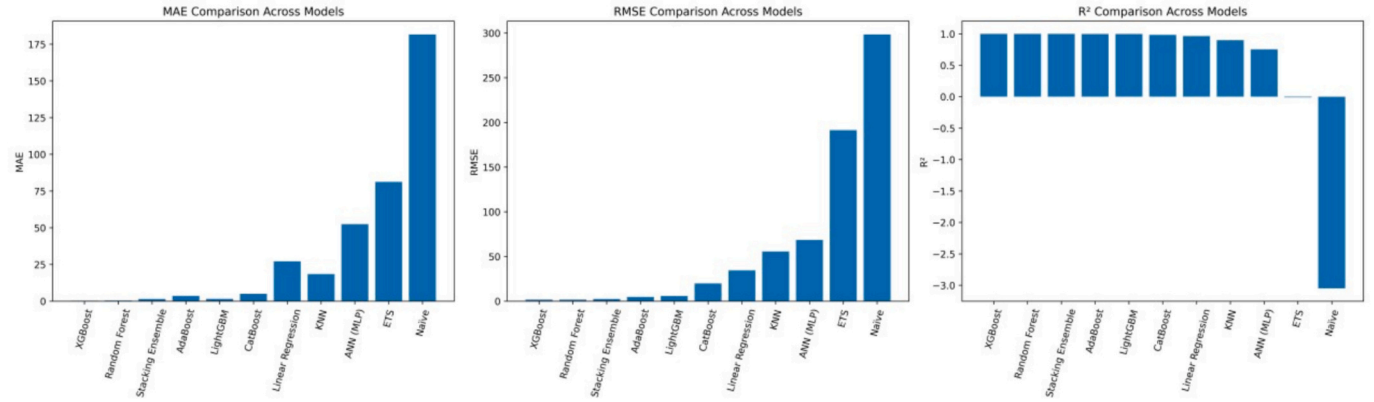


Fig. 6. Model-wise averaged metrics.

baselines, confirming that the proposed framework captures meaningful temporal structure beyond one-step prediction (Supplementary Table S4).

Walk-forward validation was performed to examine the temporal stability of model performance across sequential evaluation folds using RMSE as the error metric. The results indicate that ensemble learning approaches, particularly Random Forest, LightGBM, CatBoost, and XGBoost, maintained consistently low prediction errors across most folds, demonstrating strong robustness under progressive training conditions (Fig. 7). In contrast, ETS exhibited several pronounced error spikes, notably around folds 3 and 17–18, suggesting sensitivity to temporal variation in the dataset. The ANN model also showed moderate fluctuations, reflecting reduced stability compared with tree-based methods. Overall, the walk-forward analysis confirms that ensemble-based regression models provide more reliable and temporally consistent predictive performance than classical statistical and neural network approaches within the evaluated sequential validation framework.

Residual diagnostics were conducted for the best-performing model under the leakage-safe walk-forward setting (XGBoost). The residual series showed no systematic temporal pattern, and the residual ACF/

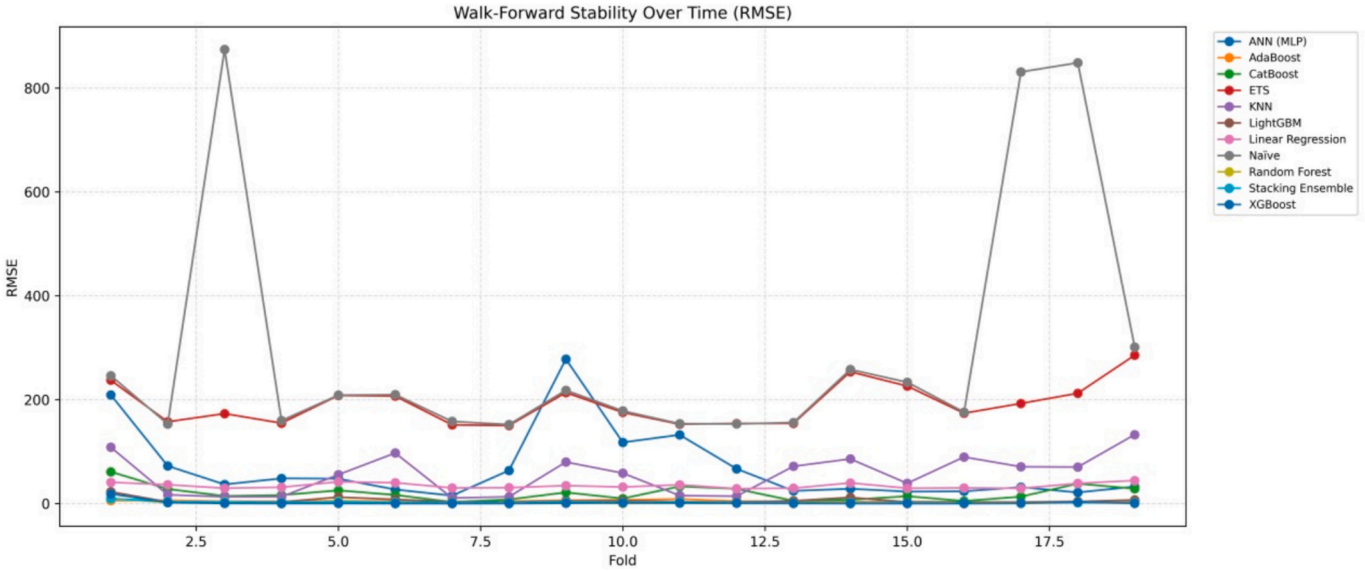


Fig. 7. Walk-Forward Stability over time (RMSE).

PACF did not exhibit persistent significant spikes, indicating that the model captured the dominant temporal structure in the data (Fig. 8). Consistently, Ljung–Box tests failed to reject the null hypothesis of no residual autocorrelation at the examined lags (Supplementary Table S2), supporting that reported forecasting skill is not primarily driven by remaining serial dependence.

The scatter plot comparing observed and predicted TDS values illustrates the strong agreement achieved by the best-performing ensemble model under the leakage-safe walk-forward evaluation. Predictions closely follow the 1:1 reference line across the dominant operational range (≈1000–1100 mg/L), indicating accurate short-term forecasting under stable conditions. Minor deviations are observed during abrupt downward spikes, which are attributed to transient sensor artifacts rather than persistent model bias. The observed–predicted agreement plot for the best-performing ensemble model (XGBoost) demonstrates strong correspondence between forecasts and measured TDS values across the dominant operational range (Fig. 9). Minor deviations are primarily associated with abrupt downward spikes attributed to transient sensor artifacts rather than persistent model bias.

The SHAP summary (beeswarm) plot provides a global overview of how individual features influence the model's predictions across all samples. Consistent with previous SHAP analyses, EC exhibits the strongest and most variable SHAP values, spanning a wide horizontal range and dominating the left side of the plot, with impacts reaching beyond −1100 SHAP units for certain instances (Fig. 10A). This demonstrates that conductivity is not only the most influential predictor on average but also the primary driver of both positive and negative deviations in TDS forecasts. In comparison, ECmV, pH, and pHmV display much smaller SHAP spreads, indicating limited influence on the global prediction landscape. The remaining features, including TDS lag values, temperature-derived variables, and pH-related statistics, cluster tightly around zero, confirming their negligible contribution across the dataset. The color distribution, where high EC values (red points) consistently create large positive contributions and low EC values (blue points) generate negative impacts, reinforces the strong monotonic relationship between conductivity and TDS.

The SHAP waterfall plot provides a complementary local, instance-level explanation by decomposing the prediction for a specific sample. Starting from the model's expected value (1045.17 mg/L), EC contributes the largest positive shift (+74.9 units) (Fig. 10B), driving the prediction upward to the final model output of 1122.76 mg/L. Secondary positive contributors, such as ECmV (+2.94) and pH (+0.94), have comparatively minor effects, while features such as pHmV (−0.92), TDS_lag_5 (−0.23), and WaterTLQE (−0.18) introduce small downward adjustments. The remaining variables collectively account for near-zero influence, as indicated by their minimal SHAP values. The structure of the waterfall plot clearly illustrates that the final prediction is shaped almost entirely by EC, with only marginal refinements from other physicochemical and lagged features. This localized explanation aligns closely with the global SHAP distribution, confirming that the model's behavior is highly consistent across both global and instance-level perspectives.

Additional residual distribution diagnostics for all models are provided in the Supplementary Material (Fig. S4).

Overall, the SHAP-based analysis confirms that TDS forecasts are predominantly driven by ionic-strength-related signals, consistent across both global and local explanations. The dominance of conductivity-related information, together with the negligible contribution of temperature and pH, explains the strong short-term predictability observed under stable aquaculture conditions. Combined with leakage-safe walk-forward validation and residual diagnostics, these results indicate that the proposed framework captures genuine time-forward predictive structure rather than artifacts of contemporaneous redundancy or serial dependence.

4. Discussion

From an operational perspective, accurate short-term forecasting of TDS enables early warning of ionic-strength excursions that may precede physiological stress in cultured organisms or reduced biofilter performance. Such early-warning capability can support timely management actions, including water exchange, dilution, or closer system inspection, before critical ecological thresholds are exceeded.

The predictive performance achieved in this study exceeds or matches that reported in most recent aquaculture water-quality forecasting research, particularly with respect to the R^2 and error magnitude. In IoT-enabled aquaculture systems, reported R^2 values for time-series forecasting of key water quality parameters typically range between 0.80 and 0.95, even when advanced ensemble or deep learning models are employed. For example, Shete et al. (2025) reported R^2 values of 0.84–0.86 for dissolved oxygen prediction and approximately 0.91 for water quality index forecasting using an IoT-driven ensemble learning framework; despite using gradient boosting and XGBoost models. Similarly; Nuangpirom et al. (2025) achieved R^2 values between

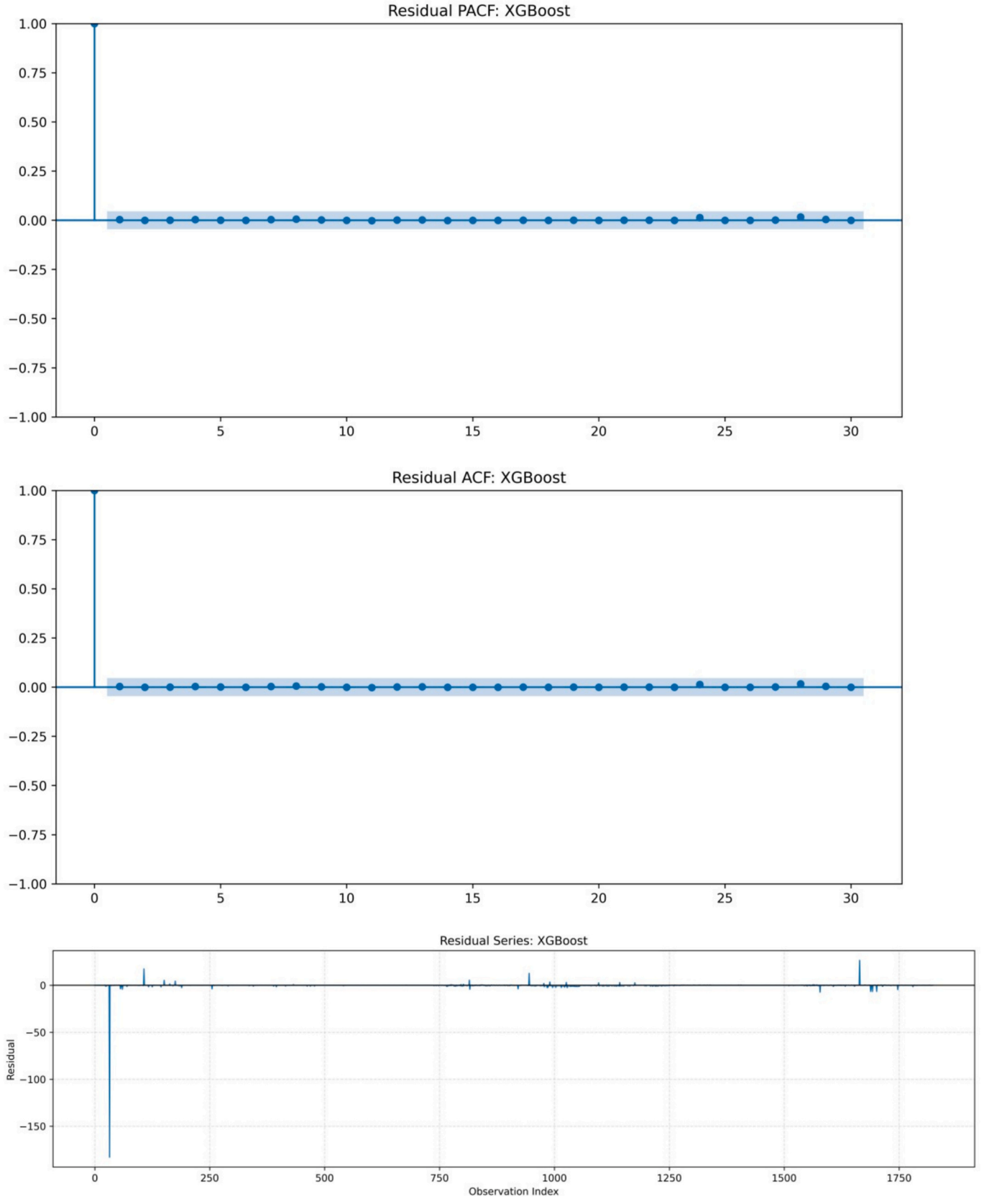


Fig. 8. (A) Forecast residual series for the best overall model (XGBoost) under the leakage-safe, lag-only specification (EC excluded); (B) Residual autocorrelation function (ACF); (C) Residual partial autocorrelation function (PACF).

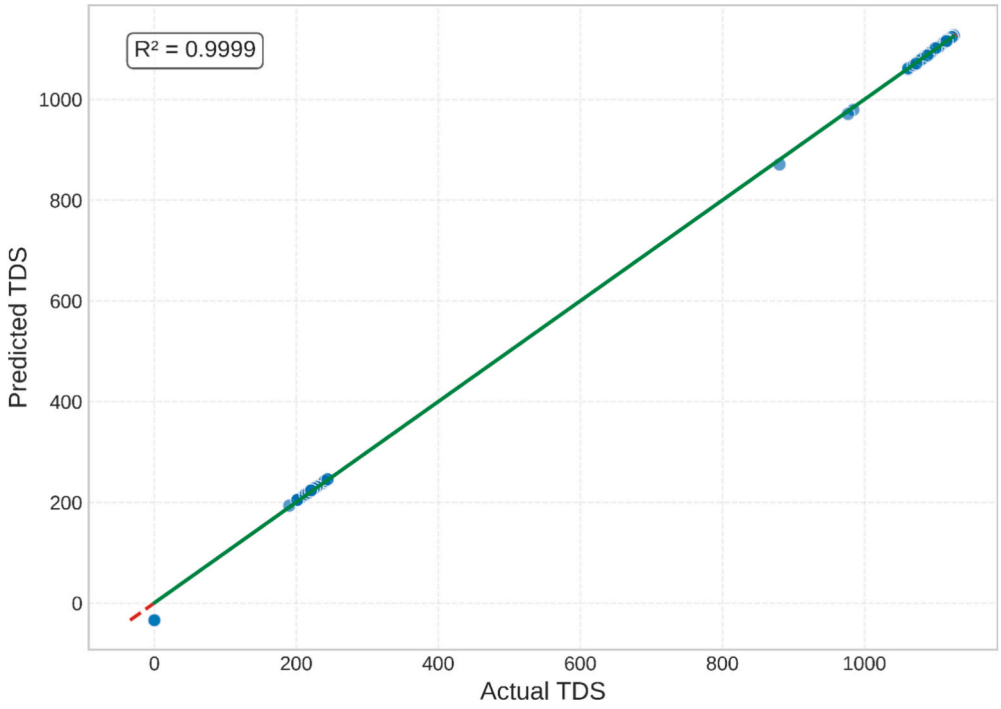


Fig. 9. Observed versus predicted TDS values for the best-performing model (XGBoost) under leakage-safe walk-forward validation. The close alignment with the 1:1 reference line across the dominant operational range indicates strong short-term forecasting accuracy.

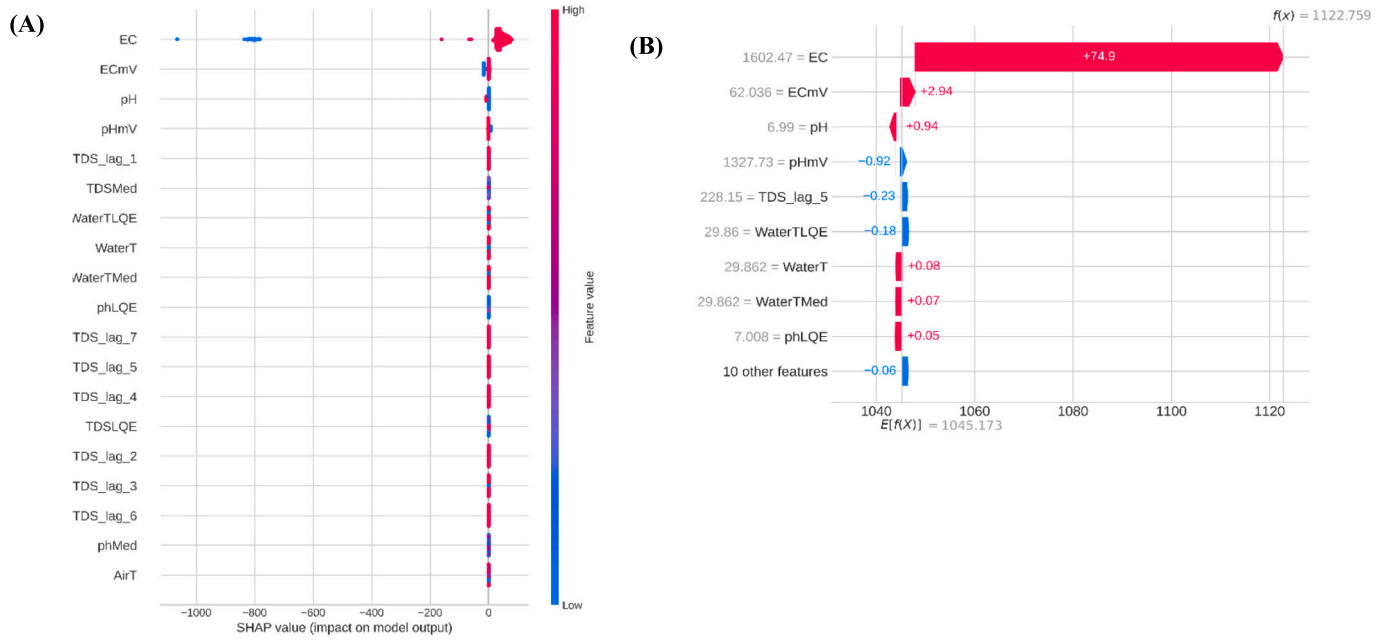


Fig. 10. Feature Contribution Analysis Using SHAP: (A) Global Beeswarm Distribution, (B) Local Waterfall Decomposition.

0.65 and 0.80 when deploying Random Forest and Multiple Linear Regression models on low-cost edge devices for real-time aquaculture monitoring, highlighting the trade-off between computational efficiency and predictive accuracy.

The exceptionally strong forecasting performance observed in this study is primarily attributable to the high temporal stability of TDS and its strong autocorrelation structure within the monitored aquaculture system. Under a leakage-safe, lag-only specification, tree-based ensemble models, particularly XGBoost, achieved the lowest prediction errors, demonstrating superior robustness to residual variability compared with linear and neural approaches. This finding indicates that, once contemporaneous redundancy is removed, nonlinear ensemble methods better capture subtle temporal patterns even in relatively stable environments.

This level of accuracy is substantially higher than that reported in most aquaculture-focused forecasting studies and even exceeds many results obtained using complex nonlinear architectures (Hridoy et al., 2026a; Kaur et al., 2023; Palaiokostas, 2021). For instance, Ewusi et al. (2021) modeled TDS in surface and groundwater systems using Gaussian Process Regression and neural networks, reporting average RMSE values

of 7.91 mg/L and R^2 values of 0.987, which are strong but still inferior to the results obtained in this study. The reduction in RMSE by more than 70% relative to these benchmarks underscores the effectiveness of lag-based linear modeling when the time series is highly autocorrelated and environmentally stable.

Deep learning approaches, while powerful in complex and highly nonlinear environments, do not consistently outperform simpler models in aquaculture settings. Shi et al. (2026) employed an optimized LSTM-RBF spatiotemporal framework for predicting dissolved oxygen and temperature, reporting notable improvements over baseline models but still observing RMSE values an order of magnitude higher than those achieved here. Similarly, Zambrano et al. (2021) demonstrated that Random Forest models could forecast water-quality parameters under sparse manual sampling conditions, yet their reported errors remained considerably larger due to limited temporal resolution and weaker autocorrelation structures.

Studies focused on surface water and river systems have occasionally reported extremely high R^2 values (0.99) for TDS prediction using hybrid or optimized ensemble models. For example, Siddiq et al. (2025) achieved $R^2 \approx 0.99$ with near-zero RMSE using PSO-optimized Random Forest models. However, such results often rely on extensive feature sets, strong preprocessing assumptions, and non-aquaculture environments, limiting direct comparability. In contrast, the present study achieves comparable or superior accuracy using a simpler, fully interpretable XGBoost model applied directly to IoT sensor data from an operational aquaculture system (Klaoudatos et al., 2025; Kol and Jairam Naik, 2025; Ramírez-Coronel et al., 2024).

Overall, the numerical comparison indicates that the exceptionally high performance observed in this study is attributable to three key factors: (i) high-frequency IoT data with minimal noise, (ii) strong linear dependence between TDS and EC, and (iii) effective lag-based time-series representation.

Fluctuations in ionic strength represent a critical yet often underappreciated driver of ecological dynamics in intensive aquatic systems (Majumdar et al., 2026). From a physiological perspective; abrupt ionic-strength excursions can impose significant osmotic stress on fish; disrupting ion exchange processes across gill membranes and increasing metabolic energy demands required for homeostasis (Huang et al., 2026). This stress can reduce growth performance; compromise immune function; and ultimately elevate mortality risks; particularly under high-density culture conditions (Hridoy et al., 2026b; Nandi et al., 2025).

In parallel, ionic strength plays a decisive role in regulating microbial processes within biofiltration units. Nitrifying bacteria, which are essential for ammonia oxidation, exhibit sensitivity to ionic imbalances that can inhibit enzymatic activity and reduce nitrification efficiency (Hossain et al., 2026; Hridoy et al., 2026c, 2026d; Lima et al., 2025). Such constraints may lead to the accumulation of toxic nitrogenous compounds, thereby exacerbating water quality deterioration and reinforcing stress feedback on cultured organisms.

At the community level, plankton assemblages demonstrate threshold-dependent responses to changes in ionic strength (Hridoy et al., 2025d; Islam et al., 2026). Even modest deviations can shift competitive interactions among phytoplankton and zooplankton, potentially triggering regime changes in community composition (Meredith et al., 2026). These shifts may influence primary productivity, trophic transfer efficiency, and overall ecosystem stability.

From an environmental informatics perspective, these multi-scale responses highlight the importance of integrating ionic-strength variability into predictive modeling frameworks (Hridoy et al., 2025e, 2025f; Tang et al., 2026). Data-driven approaches, including ML and real-time sensor analytics, offer promising avenues for detecting early-warning signals and identifying nonlinear thresholds in system behavior (Noor, 2026). Incorporating ionic strength as a key explanatory variable can enhance model robustness; improve forecasting accuracy; and support adaptive management strategies for sustainable aquaculture operations (Hridoy et al., 2026e).

Despite the strong predictive performance and interpretability achieved in this study, several limitations should be acknowledged. First, the analysis is based on data collected from a single IoT-enabled aquaculture system characterized by relatively stable physicochemical conditions. The high temporal stability of TDS, pH, and temperature contributes to the dominance of linear relationships and may limit the generalizability of the findings to aquaculture environments that experience stronger seasonal variability, abrupt disturbances, or heterogeneous management practices. Systems exposed to rainfall events, water exchange operations, algal blooms, or feeding shocks may exhibit nonlinear dynamics that were not present in the current dataset (Baydaroğlu, 2025).

Second, the forecasting task focused on short-term, one-step-ahead prediction using lag-based features. While this approach is well suited for real-time monitoring and early warning systems, it does not capture longer-horizon dependencies or cumulative effects that may influence water quality over extended periods (Liu et al., 2025a, 2025b). Multi-step forecasting can introduce error propagation and may require different modeling strategies, particularly in environments with delayed ecological responses.

Though multiple ML models were evaluated, the feature space was largely restricted to sensor-derived physicochemical variables. External drivers such as feeding rate, stocking density, water exchange volume, weather conditions, and management interventions were not available and therefore not included in the modeling framework (Godde et al., 2019; Hridoy et al., 2026f). The absence of these exogenous variables may limit the model's ability to explain rare deviations or structural changes in water quality dynamics.

Although classical statistical benchmarks were partly addressed through Naïve and ETS baselines, ARIMA/SARIMA models were not implemented in the present revision. ARIMA/SARIMA benchmarking will be considered in future work as an additional sensitivity analysis under the same rolling-origin evaluation, particularly for assessing longer-horizon and seasonal components where applicable.

Finally, while SHAP-based explainability provided strong insights into feature contributions, the interpretability analysis was primarily conducted on the best-performing XGBoost model. Although this choice aligns with the study's emphasis on transparency, explainability results from more complex nonlinear models could further enrich understanding of potential interaction effects, even if their predictive accuracy is lower.

The relatively narrow dynamic range and temporal stability of TDS in the present dataset contribute to the high predictability observed and may limit direct generalization to aquaculture systems subject to stronger seasonal forcing or abrupt management disturbances.

Importantly, this performance is observed consistently across walk-forward folds, indicating stability rather than split-specific optimism.

Future research should aim to validate the proposed interpretable forecasting framework across multiple aquaculture systems with diverse species, management intensities, and environmental conditions. Cross-site and cross-seasonal validation would help assess the robustness of linear dominance and identify conditions under which nonlinear or hybrid models become necessary.

Extending the modeling approach to multi-step and long-horizon forecasting represents an important next step. Incorporating recursive or direct multi-output forecasting strategies could provide actionable insights for medium-term planning, such as feed scheduling, water exchange optimization, and risk mitigation. Hybrid approaches that combine linear models for short-term stability with nonlinear models for long-term trend detection may offer a balanced solution.

Future studies should also integrate exogenous and operational variables, including meteorological data, feeding regimes, biomass estimates, and management actions. The inclusion of these drivers would enable causal analysis and improve the model's ability to anticipate regime shifts rather than merely extrapolate historical patterns. Such integration is particularly relevant for developing decision-support

systems that move beyond monitoring toward adaptive control.

From a methodological perspective, further exploration of explainable ensemble and hybrid models is warranted. While XGBoost proved optimal in this study, explainable boosting machines, generalized additive models, or constrained neural networks could capture mild nonlinearities while preserving interpretability. Combining SHAP with complementary explainability techniques may also enhance trust and usability for aquaculture practitioners.

5. Conclusion

This study developed an interpretable, leakage-safe machine-learning framework for short-term forecasting of aquaculture water quality, with a focus on TDS, using high-frequency sensor time series. Evaluation under an expanding-window walk-forward protocol demonstrated that tree-based ensemble models, particularly XGBoost, achieved the lowest forecasting errors under a lag-only specification that excluded redundant conductivity predictors.

SHAP-based interpretation indicated that TDS dynamics are dominated by ionic-strength-related signals, while temperature and pH exert comparatively minor influence. These results highlight the importance of combining leakage-safe validation, interpretable modeling, and ensemble learning to support reliable early-warning and decision-support systems in aquaculture. Future work should extend this framework to multi-step forecasting horizons, incorporate exogenous management drivers, and validate performance across more variable farming environments.

Clinical trial number

Internet of Things (IoT) devices and subsequent analysis using deep learning techniques. As such, clinical trial procedures do not apply to our work.

Software usage

Artificial intelligence (AI) software, including machine learning libraries such as TensorFlow, was utilized for data analysis and modeling in this study. While no AI tools were used to write or prepare the manuscript text, AI-assisted tools were employed solely to support ideation and conceptual thinking during the research process.

Third-party involvement

No persons or third-party services were involved in the research or manuscript preparation who are not listed as an author or acknowledged in the manuscript.

CRediT authorship contribution statement

Md. Abdullah Al Mamun Hridoy: Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization, Writing – review & editing, Writing – original draft. **Matteo Bodini:** Validation, Resources, Investigation, Funding acquisition, Formal analysis, Data curation, Writing – review & editing. **Munshaibur Rahman Mahin:** Formal analysis, Data curation. **Petra Schneider:** Formal analysis, Writing – review & editing. **Paolo Pastorino:** Formal analysis, Writing – review & editing. **Chiara Bordin:** Formal analysis, Writing – review & editing. **Md. Abdullah Al Mamun:** Formal analysis, Writing – review & editing. **Leonardo Goliatt:** Formal analysis, Writing – review & editing.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare no conflicts of interest. No funding, commercial, or institutional pressures influenced the design, execution, or reporting of this study. All affiliations listed were included solely for administrative and identification purposes.

Acknowledgment

The author expresses sincere gratitude to Mendeley Data Repository for providing a platform to share and access the dataset, which greatly supported the development of this research.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2026.103804.

Data availability

The datasets generated and analyzed during the current study are publicly available in the Google Drive repository:

https://drive.google.com/drive/folders/1GsTNzV5bd8u5d4DVAkH9CikbPxTNpxnS?usp=drive_link

The source code used for modeling and analysis is openly accessible via GitHub:

https://github.com/MHridoy-123/Time-series-forecasting-of-aquaculture-water-quality-

All materials are provided to support the transparency and reproducibility of the study.

References

Ahmad, T., Chen, H., Huang, R., Yabin, G., Wang, J., Shair, J., Kazim, M., 2018. Supervised based machine learning models for short, medium and long-term energy prediction in distinct building environment. Energy 158, 17–32.

Al Mamun Hridoy, M.A., Paul, P.B., Masood, A., 2025. Monitoring flood-prone urban areas of Sylhet, Bangladesh through water quality remote sensing. Discov. Environ. 3 (1), 125.

Amulejoye, F.D., Olusola, S.E., 2026. Policy and regulation on water quality and management. In: Water Remediation Methods and Wastewater Treatment. Elsevier, pp. 21–35. https://doi.org/10.1016/B978-0-443-33038-4.00016-7.

Baydaroğlu, Ö., 2025. Harmful algal bloom prediction using empirical dynamic modeling. Sci. Total Environ. 959, 178185.

Belay, M.A., Blakseth, S.S., Rasheed, A., Salvo Rossi, P., 2023. Unsupervised anomaly detection for IoT-based multivariate time series: existing solutions, performance analysis and future directions. Sensors 23 (5), 2844.

Campos, D., Galvão, V., de Rezende, M.L., Braga, A., Bodini, M., Aires, U.R., Goliatt, L., 2026. Automated machine learning achieves accurate water quality prediction with reduced parameter requirements. Sci. Rep. 16, 1–25, 4431.

Elmessery, W.M., Abdallah, S.E., Oraiath, A.A.T., Espinosa, V., Abuhussein, M.F.A., Szűcs, P., Elwakeel, A.E., 2025. A deep deterministic policy gradient approach for optimizing feeding rates and water quality management in recirculating aquaculture systems. Aquac. Int. 33 (4), 253. https://doi.org/10.1007/s10499-025-01914-z.

Ewusi, A., Ahenkorah, I., Aikins, D., 2021. Modelling of total dissolved solids in water supply systems using regression and supervised machine learning approaches. Appl Water Sci 11 (2), 1–16.

Godde, C., Dizyee, K., Ash, A., Thornton, P., Sloat, L., Roura, E., Herrero, M., 2019. Climate change and variability impacts on grazing herds: insights from a system dynamics approach for semi-arid Australian rangelands. Glob. Chang. Biol. 25 (9), 3091–3109.

Gulati, K., Boddu, R.S.K., Kapila, D., Bangare, S.L., Chandnani, N., Saravanan, G., 2022. A review paper on wireless sensor network techniques in internet of things (IoT). Mater. Today Proc. 51, 161–165. https://doi.org/10.1016/j.matpr.2021.05.067.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Hussain, A., 2024. Interpreting black-box models: a review on explainable artificial intelligence. Cogn. Comput. 16 (1), 45–74.

Hossain, M.S., Lima, M.A., Jamee, M.F., Jahan, T., Ditthakit, P., Rahman, M.S., Hridoy, M.A.A.M., 2026. Assessing the genotoxic effects of emerging organic contaminants in water systems: current trends and future directions. Scientifica 2026 (1), 5677364.

Hridoy, M.A.A.M., Paul, P.B., 2024. Assessing Total Dissolved Oxygen and Electrical Conductivity Using Sentinel-2 Remote Sensing: A Multivariable Approach to Flood Detection in Flood-Prone Urban Area Sylhet.

Hridoy, M.A., Adikari, D., Shahriar, F., Abu, M., 2021. Opportunities and strategies to achieve potential growth of fish farming in north-East Bangladesh. J. Livestock Sci. 15, 125–135.

Hridoy, M.A.A.M., Bordin, C., Masood, A., Masood, K., 2025a. Predictive modelling of aquaculture water quality using IoT and advanced machine learning algorithms. Results Chem. 16, 102456. https://doi.org/10.1016/j.rechem.2025.102456.

Hridoy, M.A.A.M., Shawkat, A.I., Bordin, C., Acharjee, M.R., Masood, A., Baki, A.O., Al Mamun, M.A., 2025b. Advanced machine learning models for accurate water quality classification and WQI prediction: implications for aquatic disease risk management. Sci. Total Environ. 1008, 180965. https://doi.org/10.1016/j.scitotenv.2025.180965.

Hridoy, M.D., Bordin, C., Morshed, M.M., Hosen, M.T., Sagor, M.F.H., Shoeb, S., Pastorino, P., 2025c. Machine-Learning Assessment of Colloidal Mechanisms Regulating Organic Carbon and Trace Metals in the Environment.

Hridoy, M.A.A.M., Ditthakit, P., Hosen, A., 2025d. Machine learning–driven prediction of heavy metal pollution for aquatic health surveillance and disease control. In: Machine Learning–Driven Prediction of Heavy Metal Pollution for Aquatic Health Surveillance and Disease Control. ISCIENCE-D-25-19889.

Hridoy, A.A.M., Neogi, S., Ujjaman, R., Hasan, M., 2025e. Water quality interactions and their synergistic effects on aquaculture performance in Bangladesh: A critical review. Results Chem. 16, 102306.

Hridoy, M.A.A.M., Paul, P.B., Hashar, M.N., Tabaschum, T., Ujjaman, R., Khan, M.S., Shakil, M.A.H., 2025f. Seasonal variations in water quality and microplastic contamination in the Surma River, Bangladesh: implications for aquatic health and human safety. Environ. Qual. Manag. 34 (3), e70055.

Hridoy, M.A.A.M., Pastorino, P., Bordin, C., Bodini, M., Dhar, N., Schneider, P., Maulud, K.N.A., 2026a. Operational agricultural water management in river-based aquaculture: a machine-learning approach to predict dissolved oxygen in the Halda River, Bangladesh. Aquac. Eng. 113, 1–13, 102693.

Hridoy, M.A.A.M., Bordin, C., Islam, M.H., Tagri, F.Z., Lima, M.A., Baki, A.O., David, G. S., 2026b. Optimizing environmental decisions on plastics as vectors of chemical pollutants in marine environments of Southeast Asia, South Asia, and East Asia: a comprehensive systematic review. J. Sea Res. 211, 1–27, 102692.

Hridoy, M.A.A.M., Sagor, P.S., Akter, P., Islam, M.H., Bordin, C., Islam, S.S., Ditthakit, P., 2026c. Decision optimization for inorganic contaminants in water systems using ecological informatics for sustainable management. Case Stud. Chem. Environ. Eng. 13, 1–17, 101362.

Hridoy, M.A.A.M., Pastorino, P., Bordin, C., Bodini, M., Jamee, M.F., Farabi, A., Schneider, P., 2026d. Wastewater-based epidemiology of hazardous organics: multi-biomarker insights and intelligent monitoring technologies. J. Hazard. Mater. Organ., 100022 https://doi.org/10.1016/j.hazmo.2026.100022.

Hridoy, M.A.A.M., Bordin, C., Pastorino, P., Maulud, K.N.A., Pathan, M.M., Baki, A.O., 2026e. Neural network-based prediction of water contamination dynamics in a rapidly developing tropical region. Case Stud. Chem. Environ. Eng. 13, 1–16, 101342.

Hridoy, M.A.A.M., Tagri, F.Z., Bordin, C., Islam, M.H., Baki, A.O., Pathan, M.M., Pastorino, P., 2026f. Occurrence and human health risks of microplastics in the Bay of Bengal using *Perna viridis* as sentinel species. Chem. Eng. J. Adv. 25, 1–16, 101039.

Huang, D., Cai, S., Hou, Y., Qin, H., Deng, Y., Li, Z., 2026. Integrated Analysis of Histophysiological Responses and Transcriptome–Metabolome Mechanisms in Coelomactra antiquata Under Ammonia Nitrogen Stress. Animals 16 (2), 192.

Islam, S.S., Mahato, S., Midya, S., 2026. Assessing water parameters and spatial patterns of zooplankton distribution in relation to the water quality index: indicators of aquatic health and ecosystem safety. Proc. Indian Natl. Sci. Acad. 1–18.

Kaur, G., Adhikari, N., Krishnapriya, S., Wawale, S.G., Malik, R.Q., Zamani, A.S., Osei-Owusu, J., 2023. Recent advancements in deep learning frameworks for precision fish farming opportunities, challenges, and applications. J. Food Qual. 2023 (1), 4399512.

Klaoudatos, D., Theocharis, A., Vardaki, C., Pachi, E., Politikos, D., Conides, A., 2025. Aspects of biology and machine learning for age prediction in the large-eye Dentex Dentex macrophthalmus (Bloch, 1791). Fishes 10 (10), 500.

Kol, B., Jairam Naik, K., 2025. Explainable deep reinforcement learning with BIGRU-A3C for early mycobacteriosis prediction in smart aquaculture. Aquac. Int. 33 (6), 541.

Lima, M.A., Islam, M.H., Neogi, S., Nasrin, K., Sen, A., Masood, A., Hridoy, Al Mamun, M. A., 2025. Recent advances in biochar technology for aquatic pollution control: a critical review of applications, barriers, and future opportunities. Discov. Sustain. 6 (1), 980.

Liu, T., He, M., Che, Z., Liu, S., Xu, L., 2025a. A hybrid enhanced optimization architecture-based model for long-term water temperature prediction in aquaculture. Aquac. Int. 33 (4), 276.

Liu, Y., Zheng, H., Zhao, J., 2025b. Enhanced water quality prediction by LSTM and graph attention network (L-GAT): an analytical study of the Pearl River Basin. Water Res. X 28, 1–14, 100383.

Majumdar, A., Moulick, D., Dey, A., Das, D., Ghosh, S., Majumder, S., Roychowdhury, T., 2026. Micro-scale microbial dynamics at the soil–water interface: biofilm architecture, non-linear response, and emerging methodological frontiers. Water 18 (6), 658.

Meredith, W., Perujo, N., Antón-Pardo, M., Romaní, A.M., Boix, D., Compte, J., Menció, A., 2026. Planktonic response to pulse or continuous inorganic nutrient inputs: temporal variations and monitoring implications. Aquat. Sci. 88 (2), 68.

Nandi, S.K., Hossain, A., Nasren, S., Kabir, M.A., Mamun, M.A.A., 2025. Effects of immediate oxygen supplementation (sodium carbonate and hydrogen peroxide) on water quality parameters, Behavioural responses and survival of Puntius sophore fingerlings. Vet. Med. Sci. 11 (6), e70624.

Neethirajan, S., 2024. Artificial intelligence and sensor innovations: enhancing livestock welfare with a human-centric approach. Hum.-Centric Intell. Syst. 4 (1), 77–92.

Nogueira, L.S., de Carvalho, M.A., Santos, B.D.O., Yonaba, R., Bamal, A., Uddin, M.G., Goliatt, L., 2026. A comparative study of ensemble and non-ensemble machine learning methods for predicting river pollution index. Eco. Inform., 103617 https://doi.org/10.1016/j.ecoinf.2026.103617.

Noor, U.A., 2026. Machine learning innovations in revolutionizing earthquake engineering: A review: UA Noor. Archiv. Computation. Methods Eng. 33 (1), 687–743.

Nuangpirom, P., Pitjamit, S., Jaikampan, V., Peerakam, C., Nakkiew, W., Jewpanya, P., 2025. Machine learning on low-cost edge devices for real-time water quality prediction in Tilapia aquaculture. Sensors 25 (19), 6159.

Palaiokostas, C., 2021. Predicting for disease resistance in aquaculture species using machine learning models. Aquacult. Rep. 20, 100660.

Ramírez-Coronel, F.J., Rodríguez-Elías, O.M., Esquer-Miranda, E., Pérez-Patricio, M., Pérez-Báez, A.J., Hinojosa-Palafox, E.A., 2024. Non-invasive fish biometrics for enhancing precision and understanding of aquaculture farming through statistical morphology analysis and machine learning. Animals 14 (13), 1850.

Saidu, M., 2025. Contributions of fisheries and aquaculture to food security in Africa. In: Food Security, Nutrition and Sustainability through Aquaculture Technologies. Springer Nature Switzerland, Cham, pp. 493–502. https://doi.org/10.1007/978-3-031-75830-0_28.

Shete, R.P., Shekhar C, A., Mahajan, Y.V., Bongale, A.M., Dharrao, D., 2025. IoT-driven ensemble machine learning model for accurate dissolved oxygen prediction in aquaculture. Discov. Internet Things 5 (1), 94.

Shi, B., Jin, X., Hu, Y., Jiang, J., Sun, Y., 2026. Precision prediction of aquaculture water quality: a spatiotemporal model integrating optimized-LSTM and radial basis function neural networks. PeerJ Comput. Sci. 12, e3515.

Siddiq, B., Javed, M.F., Aldrees, A., 2025. Machine learning-driven surface water quality prediction: an intuitive GUI solution for forecasting TDS and DO levels. Water Qual. Res. J. 60 (4), 514–546.

Tang, J., Zhang, F., He, X., Nie, S., Ma, X., Ahmed, Z., Oke, S.A., 2026. Decoupling hydrodynamic drivers of suspended sediment in hypersaline lakes: A physics-informed machine learning approach. Int. J. Appl. Earth Obs. Geoinf. 148, 105222.

Tina, F.W., Afsarimanesh, N., Nag, A., Alahi, M.E.E., 2025. Integrating AIoT technologies in aquaculture: a systematic review. Fut. Internet 17 (5), 199.

Zambrano, A.F., Giraldo, L.F., Quimbayo, J., Medina, B., Castillo, E., 2021. Machine learning for manually-measured water quality prediction in fish farming. PLoS One 16 (8), e0256380.