# An Interpretable Machine Learning Framework for River Water Hardness Classification

Aditya Rajiv Ratnam[1]

[1]National Academy for Learning, Bengaluru, Karnataka, India

## Abstract

Determining water hardness in the field currently relies on laboratory Ethylenediaminetetraacetic acid (EDTA) titration or single-proxy electrical conductivity (EC) measurements, both of which have significant limitations. This study investigates whether routine in situ physicochemical parameters can accurately classify river water hardness without laboratory analysis. Using samples from a river basin in Argentina, four ML classifiers were evaluated against a trivial majority-class classifier (71.4% accuracy) and an EC threshold baseline (75% accuracy). A Random Forest model achieved 95.45% accuracy, a 20.45 percentage point improvement over the EC baseline. Five complementary feature importance and model interpretability approaches (Gini importance, permutation importance, SHapley Additive exPlanations (SHAP), feature ablation, and correlation analysis) show that the model captures physicochemical relationships pertaining to ionic interference, hydrological dilution, and temperature-dependent variations in conductivity. A minimal three-sensor suite comprising EC, Total Suspended Solids (TSS), and sample temperature matched full-model performance, demonstrating that accurate hardness classification is feasible with a low-cost, reagent-free field system. These findings support a practical pathway toward simpler and more sustainable water quality monitoring, in alignment with green chemistry principles of waste prevention and reduced chemical use.

**Keywords:** *water hardness classification, machine learning, random forest, EDTA titration, water quality monitoring, physicochemical parameters, feature selection, sensor reduction, green chemistry, electrical conductivity*

## 1. Introduction

### 1.1 Water Hardness and the Field Monitoring Challenge

Freshwater quality monitoring is among the most pressing environmental challenges of the 21st century. Rivers serve simultaneously as sources of drinking water, irrigation supply, and biodiversity, making accurate and cost-effective chemical assessment a matter of public health and ecological concern. Water hardness is a measure of the concentration of divalent cations, principally calcium ($Ca^{2+}$) and magnesium ($Mg^{2+}$), expressed as milligrams of equivalent calcium carbonate per liter (mg $CaCO_3$/L). Hard water accelerates scale formation in pipes and boilers, affects the bioavailability of trace metals in aquatic systems, and influences the palatability and lathering properties of domestic water (World Health Organization, 2011).

The chemistry of hardness in natural waters is governed by the carbonate equilibrium system. Dissolved $CO_2$ hydrates to form carbonic acid ($H_2CO_3$), which dissociates sequentially:

$$CO_2(aq) + H_2O \rightleftharpoons H_2CO_3 \rightleftharpoons H^+ + HCO_3^- \rightleftharpoons CO_3^{2-} + 2H^+. \quad (eq1)$$

In the pH range of 7–9 typical of river water, bicarbonate buffering dominates. Calcium and magnesium carbonates are sparingly soluble; lower pH and higher $pCO_2$ promote dissolution:

$$CaCO_3(s) + CO_2(aq) + H_2O \rightleftharpoons Ca^{2+} + 2HCO_3^-. \quad (eq2)$$

An analogous reaction occurs for $Mg^{2+}$. As these equilibrium processes depend on carbonate mineral availability, water hardness is fundamentally controlled by the river basin's geological composition. Rivers passing through limestone, dolomite, or chalk substrates accumulate $Ca^{2+}$ and $Mg^{2+}$ from mineral dissolution, while those draining silicate or granite terrain carry fewer divalent cations (Hem, 1985). These geological differences produce persistent spatial patterns in hardness. Seasonal rainfall also introduces temporal variability. Together, this makes single-point sampling insufficient.

Existing approaches for measuring water hardness involve a trade-off between accuracy and practical deployability. EDTA, the reference standard, determines water hardness through complexometric titration (Baird, 2017). At pH 10, EDTA binds $Ca^{2+}$ and $Mg^{2+}$ in a 1:1 molar ratio, displacing a metal-indicator complex at the endpoint. The reaction, with an analogous equivalent for $Mg^{2+}$, is:

$$Ca^{2+} + EDTA^{4-} \rightarrow [CaEDTA]^{2-} \quad (eq3)$$

While rigorous, this procedure requires buffered solutions, indicator reagents, a burette, and trained personnel, which are impractical in remote field settings and continuous monitoring. In contrast, field sensors commonly estimate hardness indirectly through EC or Total Dissolved Solids (TDS), which track total ionic strength rather than hardness. These proxy-based methods are inherently limited by ionic interference, as monovalent species such as $Na^+$ and $Cl^-$ contribute to EC without increasing hardness, leading to systematic misclassification. Advanced analytical techniques can provide ion-specific measurements with high precision, but their cost and operational constraints limit their use in large-scale or real-time monitoring. Machine learning (ML) methods have recently been applied to water quality prediction as a cost-effective alternative (Haghiabi et al., 2018; Zhu et al., 2022).

### 1.2 Water Hardness Classification Schemes

The World Health Organization (WHO, 2011) classifies water hardness as *soft* (< 60 mg $CaCO_3$/L), *moderately hard* (60–120 mg $CaCO_3$/L), *hard* (120–180 mg $CaCO_3$/L), and *very hard* (> 180 mg $CaCO_3$/L). The source dataset employs locally-defined categories in Spanish: *blanda* (< 150 mg $CaCO_3$/L), encompassing the WHO soft and moderately hard classes, and *semidura* (≥ 150 mg $CaCO_3$/L), corresponding to the WHO hard and very hard categories.

### 1.3 Research Questions

This study addresses the following questions:

1. Can routine in situ physicochemical measurements predict hardness classification without direct titration?
2. Which ML algorithm provides the best classification performance, and what characteristics of the underlying predictive structure explain its superiority?
3. To what extent does the available dataset support reliable hardness classification within the sampled catchment?
4. Which physicochemical parameters are the most predictive for hardness classification, and can their contributions be explained in terms of known hydrochemical mechanisms?
5. Can a reduced feature set preserve classification performance, and what potential advantages does this offer for practical deployment?

## 2. Methods and Models

### 2.1 Dataset and Study Area

The dataset used in this study is available on Kaggle (Ferran, 2024), and was generated through longitudinal field sampling conducted by a research team in Argentina. Over 12 field campaigns from May to November 2023, a total of 219 water samples were collected from five locations within the anthropogenically impacted Matanza–Riachuelo river basin in Buenos Aires province. The dataset is chronologically ordered. After removing entries with missing values, 217 samples were retained. The catchment drains a mixed agricultural and urban landscape with Pampean loess deposits having significant carbonate mineral content, contributing to naturally elevated ionic loads. Laboratory analyses included EDTA titration for hardness (mg $CaCO_3$/L). Table 1 summarizes the in situ physicochemical features and their descriptive statistics.

| Parameter | Mean | SD | Min | Max | Missing |
|---|---|---|---|---|---|
| **Ambient temperature (°C)** | 18.06 | 5.52 | 10.40 | 30.50 | 0 |
| **Ambient humidity** | 0.57 | 0.19 | 0.19 | 0.87 | 0 |
| **Sample temperature (°C)** | 19.31 | 4.11 | 12.80 | 28.10 | 0 |
| **pH** | 8.03 | 0.29 | 7.20 | 8.70 | 0 |
| **EC (µS/cm)** | 1264.57 | 273.32 | 200.00 | 1710.00 | 0 |
| **TDS (mg/L)** | 624.25 | 135.54 | 140.00 | 850.00 | 0 |
| **TSS (mL sed/L)** | 49.83 | 86.21 | 0.10 | 650.00 | 6 |
| **DO (mg/L)** | 2.62 | 1.96 | 0.00 | 9.12 | 0 |
| **Level (cm)** | 36.50 | 10.40 | 10.00 | 60.00 | 39 |
| **Turbidity (NTU)** | 144.95 | 234.59 | 1.06 | 1000.00 | 1 |
| **Total $Cl^-$ (mg/L)** | 98.62 | 38.40 | 32.00 | 174.00 | 4 |
| **Hardness* (mg $CaCO_3$/L)** | 190.71 | 56.06 | 86.00 | 316.00 | 2 |

**Table 1. Descriptive Statistics of Physicochemical Features (n = 217)**

* Excluded from modeling to prevent data leakage (see Section 2.3). Level has the highest proportion of missing values, followed by TSS and Total $Cl^-$; all missing values were imputed using column medians.

### 2.2 Implementation and Code Availability

The complete pipeline described in this paper is publicly available on GitHub (Ratnam, 2026), and includes an interactive water hardness predictor that allows users to explore model predictions by adjusting input feature values through graphical sliders. It retains the original variable designations and units used in the source dataset and uses *scikit-learn* (Pedregosa et al., 2011) for all classifiers and evaluation routines.

### 2.3 Data Preprocessing and Validation Strategy

The measured hardness value (mg $CaCO_3$/L) was excluded to prevent data leakage. As the classification label is derived by thresholding this variable, including it would allow the classifier to trivially reproduce the threshold rather than learn generalizable relationships (Kaufman et al., 2012). The eleven remaining numeric features formed the model input. Missing values in TSS (6), Level (39), Turbidity (1), and Total $Cl^-$ (4) were imputed using column medians, a strategy robust to the extreme values present in turbidity, where episodic storm-runoff readings of 1,000 NTU represent ecologically valid observations rather than

errors. No outlier removal was performed. All models received the same stratified 80:20 train-test split (173 training samples and 44 test samples). These shared preprocessing steps ensure that the performance differences among models only reflect differences in learning algorithms.

Five-fold stratified CV with shuffling was employed to preserve class proportions while randomizing fold assignment. To prevent preprocessing leakage during CV, imputation and feature scaling were performed within each model's pipeline. For each CV fold, both the imputer and scaler were fitted exclusively on the training partition and then applied to the corresponding validation partition. Each model was then trained on the complete preprocessed training set and evaluated on the held-out test set.

## 2.4 Baseline Models

### **2.4.1** *The trivial semidura classifier*

The dataset exhibits a class imbalance, with approximately 71.4% of samples classified as *semidura* (n = 155) and 28.6% as *blanda* (n = 62). This distribution reflects the naturally elevated mineral content of the river system, where the majority of observations fall within the higher-conductivity *semidura* range associated with carbonate-bearing geology. This imbalance establishes a statistical baseline: a trivial classifier that always predicts *semidura* would achieve 71.4% accuracy. Any meaningful predictive model must significantly exceed this threshold, demonstrating that it has learned chemically relevant structure rather than simply reproducing class frequencies.

### **2.4.2** *The EC threshold classifier*

Electrical Conductivity (EC) measures the ability of water to conduct electric current, and is proportional to dissolved ion concentration and mobility. Because $Ca^{2+}$ and $Mg^{2+}$ are primary hardness contributors and significant conductivity carriers, EC serves as a natural single-parameter proxy for hardness in carbonate-dominated catchments, although its relationship is affected by other dissolved ions and temperature. The second baseline classifies samples using the training-set median EC value of 1,300 µS/cm, assigning *semidura* above the threshold and *blanda* at or below. The median was used as a simple, data-derived reference threshold rather than a universal hydrochemical cutoff. This rule requires no computation beyond a single field measurement and provides a reference benchmark that an ML model must exceed to justify its additional complexity.

However, monovalent ions such as $Na^{+}$ and $Cl^{-}$ elevate conductivity without increasing hardness, limiting EC as a standalone predictor. This interference is grounded in *Kohlrausch's Law of Independent Ion Migration*, which states that the total molar conductivity of an electrolyte solution equals the sum of the individual ionic contributions:

$\Lambda m = \Sigma\, \nu_i \lambda_i$ ($\lambda_i$: molar conductivity of ion $i$; $\nu_i$: its stoichiometric coefficient) *(eq4)*

As EC reflects the combined contribution of all dissolved ions, a high-conductivity sample may result from elevated $Na^{+}$ and $Cl^{-}$ inputs from agricultural runoff rather than from $Ca^{2+}$ and $Mg^{2+}$ responsible for hardness. The differing molar conductivities illustrate this interference: $Ca^{2+}$ (≈ 119 S·cm²/mol), $Mg^{2+}$ (≈ 106.0 S·cm²/mol), $Na^{+}$ (≈ 50.1 S·cm²/mol), $Cl^{-}$ (≈ 76.3 S·cm²/mol) all contribute to EC according to both their concentration and ionic mobility. EC also increases with temperature due to decreased water viscosity, following Walden's rule linking ionic mobility to solvent viscosity. This introduces a thermal artifact that the model must separate from the genuine ionic signal. These limitations motivate the

multivariate ML approach capable of distinguishing hardness-causing divalent cation patterns from other ionic contributions.

### 2.5 Machine Learning Models

Four ML classifiers were selected to span a range of decision boundary geometries and complexity levels. Logistic Regression (LR; L2 regularization, 1,000 iterations) models class log-odds as a linear combination of features, providing interpretable coefficients but requiring a linearly separable boundary. K-Nearest Neighbors (KNN; $k$ = 5, Euclidean distance) classifies each sample by majority vote among its nearest neighbors, but is scale-sensitive and degrades in high-dimensional spaces. The Support Vector Machine (SVM; Radial Basis Function (RBF) kernel, default $C$ and $\gamma$) learns a maximum-margin separating hyperplane extended to non-linear boundaries through the RBF kernel function. The Random Forest (RF; 100 trees, *random_state* = 42) constructs an ensemble of decision trees on bootstrapped subsets with random feature subsampling at each split, enabling robust modeling of non-linear relationships and interactions. LR, KNN, and SVM inputs were *z-score* normalized. RF was trained on unscaled features, as tree-based models do not require feature scaling.

### 2.6 Model Evaluation Framework

Performance was assessed on the held-out test set ($n$ = 44, stratified 80:20 split) using accuracy, per-class precision, recall, and F1-score. Macro and weighted F1-scores are also reported, with the latter accounting for the 71:29 class imbalance. The area under the Receiver Operating Characteristic (ROC) curve (AUC) measures class separability across all decision-score thresholds, providing an imbalance-robust complement to accuracy. Learning curve analysis was conducted on the best-performing classifier using the same shuffled stratified five-fold CV, preserving class proportions across folds. Model performance was evaluated across progressively larger training subsets drawn from the full dataset, to assess how it changed with increasing training-set size. A confusion matrix was computed on the held-out test set to decompose aggregate accuracy into per-class prediction counts, revealing directional bias.

### 2.7 Feature Importance and Model Interpretability Methods

Gini importance, measured as the mean decrease in node impurity, provided an initial estimate of feature importance. Permutation importance was calculated by repeatedly shuffling each feature and measuring the resulting decrease in test-set accuracy. SHAP (Lundberg & Lee, 2017) provided sample-level attribution, quantifying each feature's directional contribution to individual predictions. All three methods were applied to the best-performing classifier, supplemented by: (a) an ablation study removing individual features and measuring the change in test accuracy; (b) a Pearson inter-feature correlation matrix quantifying pairwise redundancy; and (c) a feature minimization study identifying the smallest feature subset that preserves classification accuracy.

## 3. Results

### 3.1 Baseline Performance

As discussed in Section 2.4.1, the trivial majority-class classifier always predicting *semidura* achieves 71.4% accuracy. The EC threshold classifier achieves 75.00% accuracy, which

suggests that reliance on conductivity alone misclassifies approximately one in four samples, highlighting a limitation of traditional EC-proxy-based approaches.

## 3.2 Model Comparison

### 3.2.1 *Accuracy and Weighted Metrics*

Table 2 shows test-set performance for all four classifiers. In the top tier, the RF classifier achieves 95.45% accuracy, a substantial margin over SVM (79.55%), LR (72.73%), and KNN (70.45%). Notably, both KNN and LR fall below the single-feature EC-threshold baseline (75.00%), suggesting that multivariate input without proper interaction modeling can introduce more noise than a single well-chosen chemical proxy for this dataset. The RF classifier achieves a weighted F1-score of 0.95, the highest value across all models. SVM follows with weighted F1 of 0.78, while LR (weighted F1 of 0.72) and KNN (weighted F1 of 0.71) remain near the statistical floor.

| Model | Acc. | Prec.(B) | Rec.(B) | F1(B) | Prec.(S) | Rec.(S) | F1(S) | Macro F1 | Weighted F1 |
|---|---|---|---|---|---|---|---|---|---|
| EC | 75.00% | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| KNN | 70.45% | 0.50 | 0.54 | 0.52 | 0.80 | 0.77 | 0.79 | 0.65 | 0.71 |
| LR | 72.73% | 0.55 | 0.46 | 0.50 | 0.79 | 0.84 | 0.81 | 0.66 | 0.72 |
| SVM | 79.55% | 0.75 | 0.46 | 0.57 | 0.81 | 0.94 | 0.87 | 0.72 | 0.78 |
| **RF** | **95.45%** | **1.00** | **0.85** | **0.92** | **0.94** | **1.00** | **0.97** | 0.94 | 0.95 |

**Table 2. Classification Performance on Held-Out Test Set (n = 44)**

B = *blanda*; S = *semidura*. EC Threshold is a single-feature rule; per-class metrics not reported. RF achieves the highest accuracy (95.45%) and weighted F1 (0.95), with perfect precision on *blanda* (1.00) and perfect recall on *semidura* (1.00). SVM, KNN and LR trail substantially.

### 3.2.2 *ROC Analysis and AUC*

Figure 1 shows the consolidated ROC curves for all four classifiers. The RF classifier achieves an AUC of 0.9888 on the test set, indicating strong class separability under the evaluated split. SVM follows with an AUC of 0.8635, while KNN (AUC = 0.7531) and LR (AUC = 0.7519) exhibit lower separability. The near-identical AUC scores of KNN and LR indicate similar discriminative performance across classification thresholds, with neither classifier providing a substantial advantage over the other in terms of AUC.

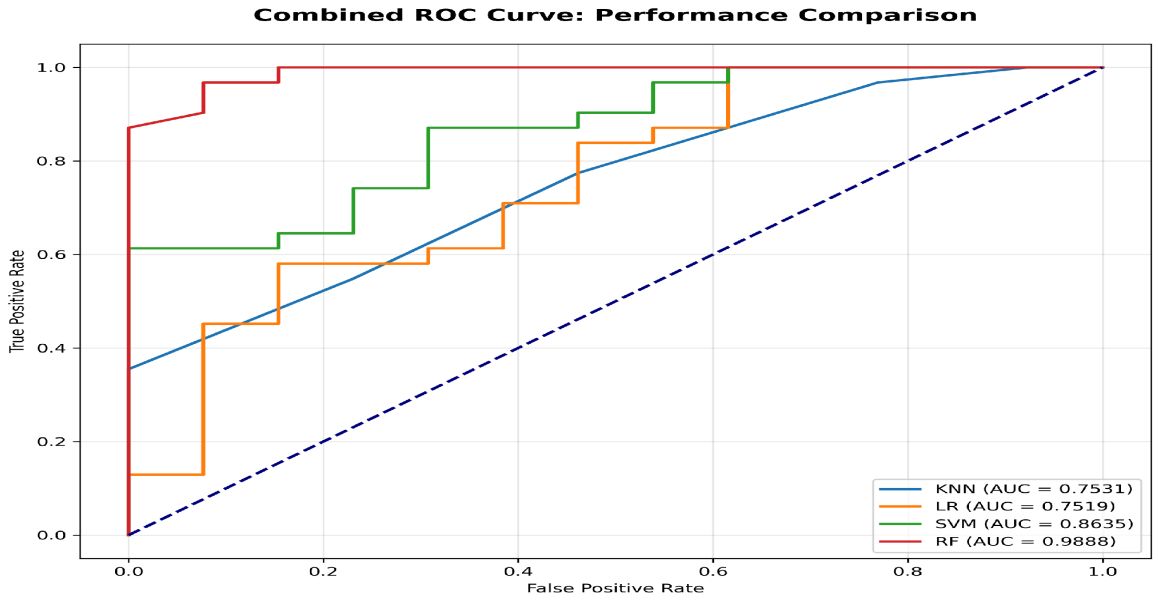


**Figure 1.** ROC curves for all classifiers (test set, n = 44). Dashed diagonal = random classifier (AUC = 0.50). RF: AUC = 0.9888. SVM: AUC = 0.8635. KNN: AUC = 0.7531. LR: AUC = 0.7519.

### 3.3 Cross-Validation (CV) Results

Figure 2 shows the CV stability for all models as box plots of fold-level accuracy scores. Boxes show the interquartile range (IQR), with whiskers extending to 1.5 times the IQR. Under a five-fold stratified CV framework, the RF classifier achieved the highest mean CV accuracy of 89.84% (± 4.80% SD). The slightly higher accuracy observed on the held-out test set (95.45%) aligns with sampling variability in the smaller test partition (n=44). A 95% Wilson score confidence interval around 95.45% (42/44 correct) gives 84.9%–98.7%, confirming the wide uncertainty inherent in small test sets and supporting the use of CV accuracy as the more conservative estimate. In comparison, SVM achieved a mean CV accuracy of 80.66% (± 3.95% SD), while KNN averaged 79.70% (± 2.83% SD) and LR averaged 78.78% (± 5.82% SD).

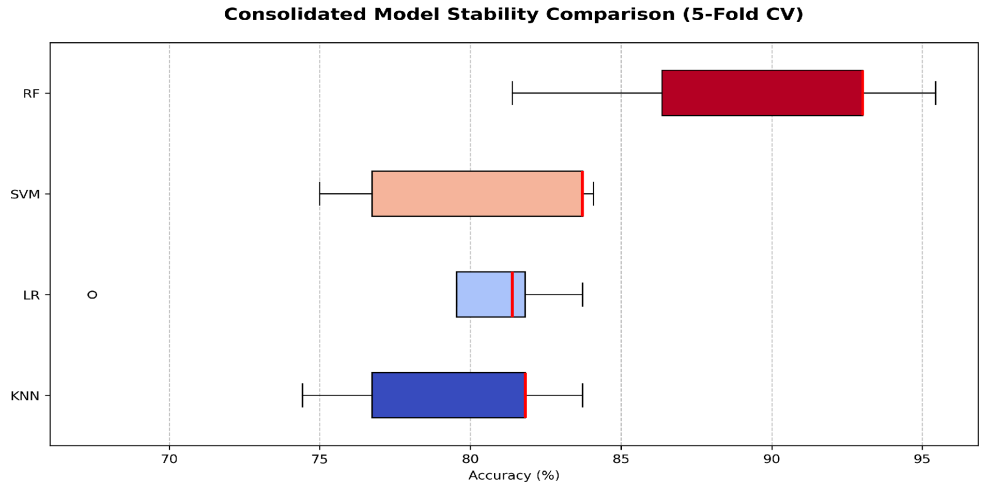


**Figure 2.** Five-fold CV accuracy distributions (box = IQR, red line = median, whiskers = 1.5·IQR, circles = outliers). RF achieves the highest mean five-fold CV accuracy (89.84% ± 4.80% SD), followed by SVM (80.66% ± 3.95% SD), KNN (79.70% ± 2.83% SD), and LR (78.78% ± 5.82% SD).

### 3.4 Learning Curve Results

Figure 3 shows the learning curve for the RF model. CV accuracy increased sharply between approximately 20 and 100 training examples and then rose more gradually as the training size approached its maximum. Training accuracy remained at 1.00 throughout, while CV accuracy remained lower. The curve had not fully plateaued at the maximum training size.

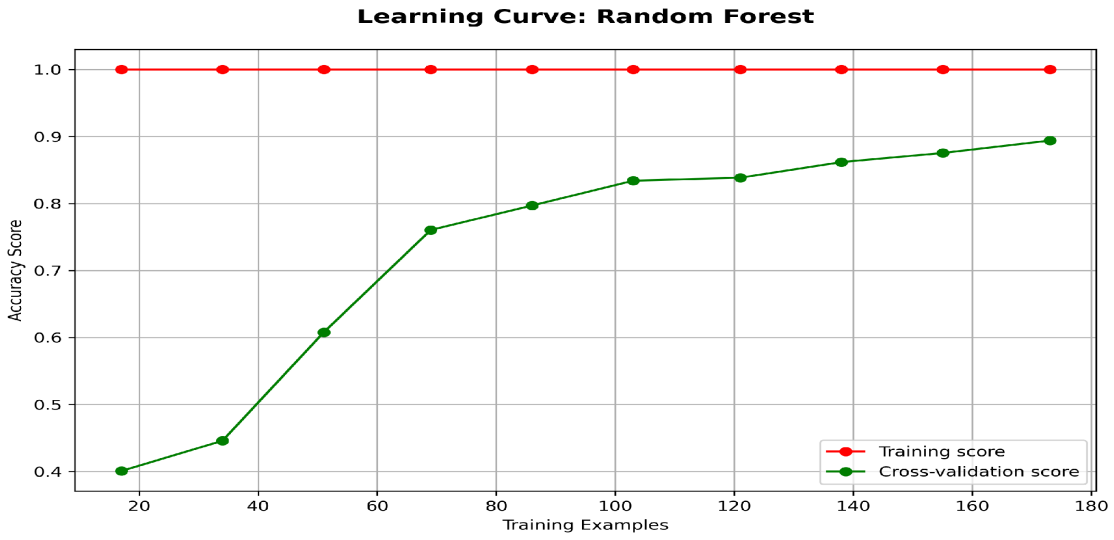


**Figure 3.** RF learning curve. Training accuracy (red) holds at 1.00 throughout; CV accuracy (green) rises from ~0.40 at n = 17 to ~0.90 at n = 173, with the rate of improvement decelerating as training size approaches its maximum. The persistent gap between training and CV curves suggests overfitting.

### 3.5 Confusion Matrix

Figure 4 presents the RF confusion matrix, showing high classification performance for both classes, although recall was higher for *semidura* than for *blanda*. Of the 44 test samples, the model correctly classified 42 instances, including 31 *semidura* and 11 *blanda* samples. Only two misclassifications were observed, both corresponding to *blanda* samples predicted as *semidura*. The matrix is asymmetric: zero *semidura* samples were misclassified as *blanda*; implications for deployment are discussed in Section 4.5.

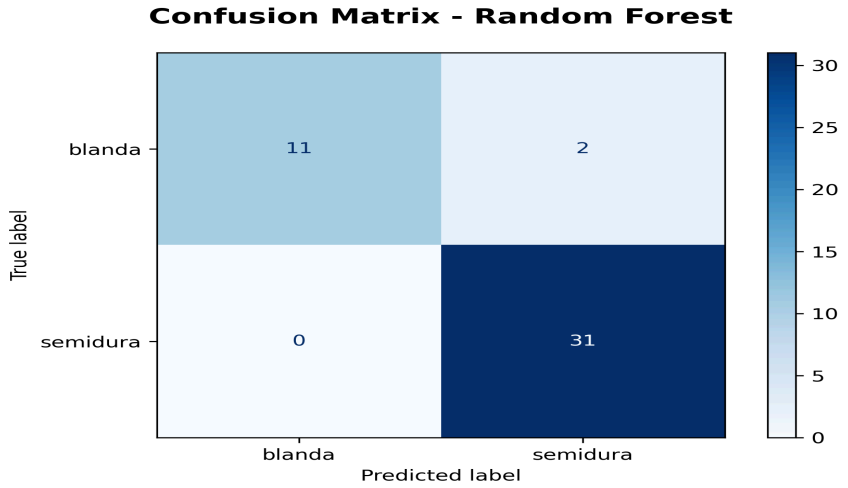


**Figure 4.** Confusion matrix for the best-performing classifier (test set, n = 44). Of 44 samples, 42 were correctly classified. Both errors were *blanda* samples predicted as *semidura*.

### 3.6 Feature Importance and Model Interpretability Results

#### 3.6.1 *Gini Feature Importance*

Gini-based feature importance, computed as the mean decrease in node impurity across all 100 trees in the RF ensemble, provides an initial estimate of the relative importance of each predictor. Each feature's score is the average amount it reduces classification uncertainty across the ensemble's 100 trees. As shown in Figure 5, turbidity emerged as the most influential feature, followed by sample temperature and EC. pH and Level ranked the lowest. Gini importance may exhibit bias toward high-cardinality features and can distribute importance unevenly when predictors are correlated, as examined further through correlation analysis in Section 3.9.

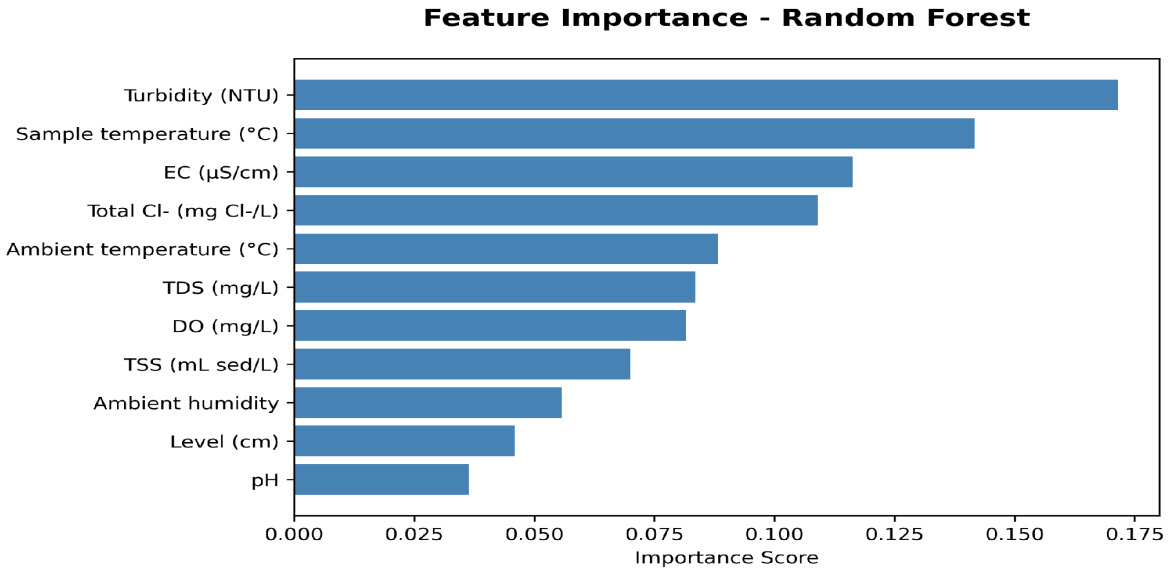


**Figure 5.** Gini feature importance scores for all eleven features (mean decrease in impurity, 100 trees). Turbidity ranks highest, followed by sample temperature and EC; pH and Level rank lowest.

### 3.6.2 *Permutation Importance*

Permutation importance was computed on the held-out test set using 10 repeats, measuring the drop in accuracy when one feature's values were randomly shuffled (Figure 6), while keeping all other features fixed. Boxes show the IQR across repeats; whiskers extend to 1.5 times the IQR. The permutation results agreed with Gini importance in identifying turbidity as a highly influential feature with the highest median drop, but several rankings differed substantially. In particular, DO rose from seventh under Gini to second under permutation, whereas sample temperature fell from second to seventh. pH and Level ranked lowest again.

Together, the contrasting rankings illustrate that Gini importance and permutation importance provide complementary but non-equivalent measures of model reliance. Because correlated predictors can substitute for one another, permutation importance may underestimate a feature's unique contribution. The substantial differences between the rankings should therefore not be interpreted as a definitive ordering of physicochemical importance.

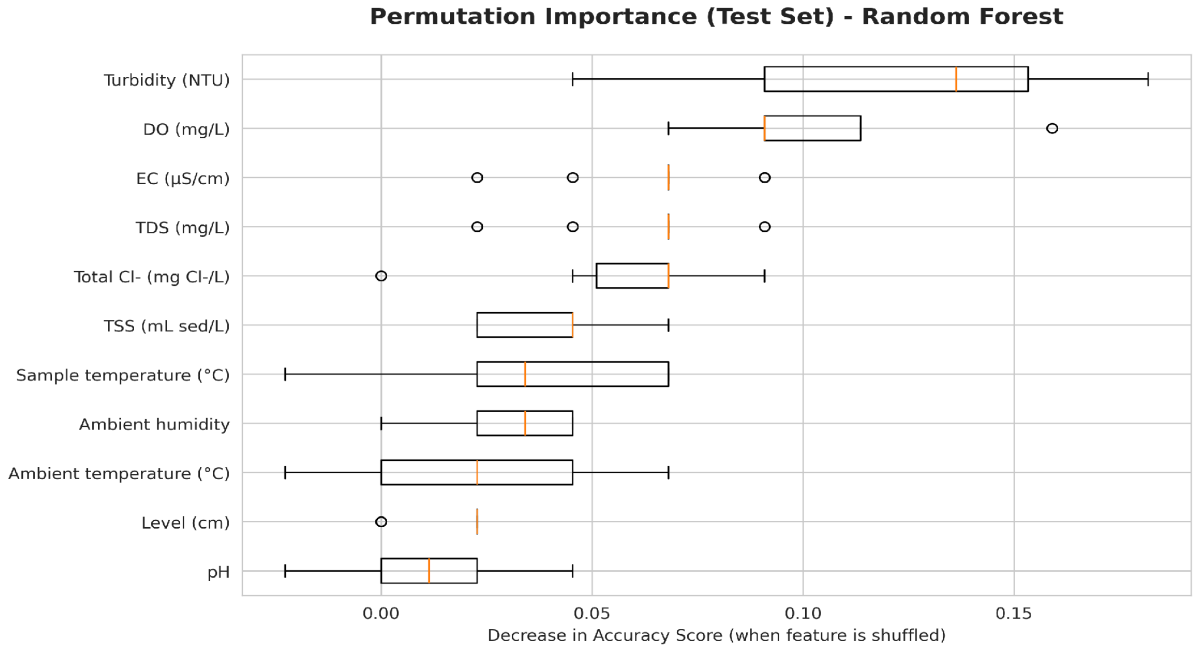


**Figure 6.** Permutation importance of RF features (10 repeats; box = IQR, orange line = median, whiskers = 1.5·IQR). Turbidity dominates; DO rises to second rank.

## 3.7 SHAP Results

To move beyond overall rankings and understand the direction and magnitude of each feature's influence on individual predictions, SHAP values were computed for the RF classifier (Figure 7). Each point represents one test sample: its horizontal position shows the SHAP value (positive values signify pushes toward *semidura*; negative values toward *blanda*), and color reflects the feature value (red = high, blue = low). SHAP ranked sample temperature, turbidity, and EC as the features with the largest mean absolute contributions to individual predictions. High sample temperature and EC are associated with positive SHAP values, pushing predictions toward the *semidura* class, while high turbidity is associated with negative SHAP values, pushing predictions toward the *blanda* class. This pattern is interpreted in physicochemical terms in Section 4.3.

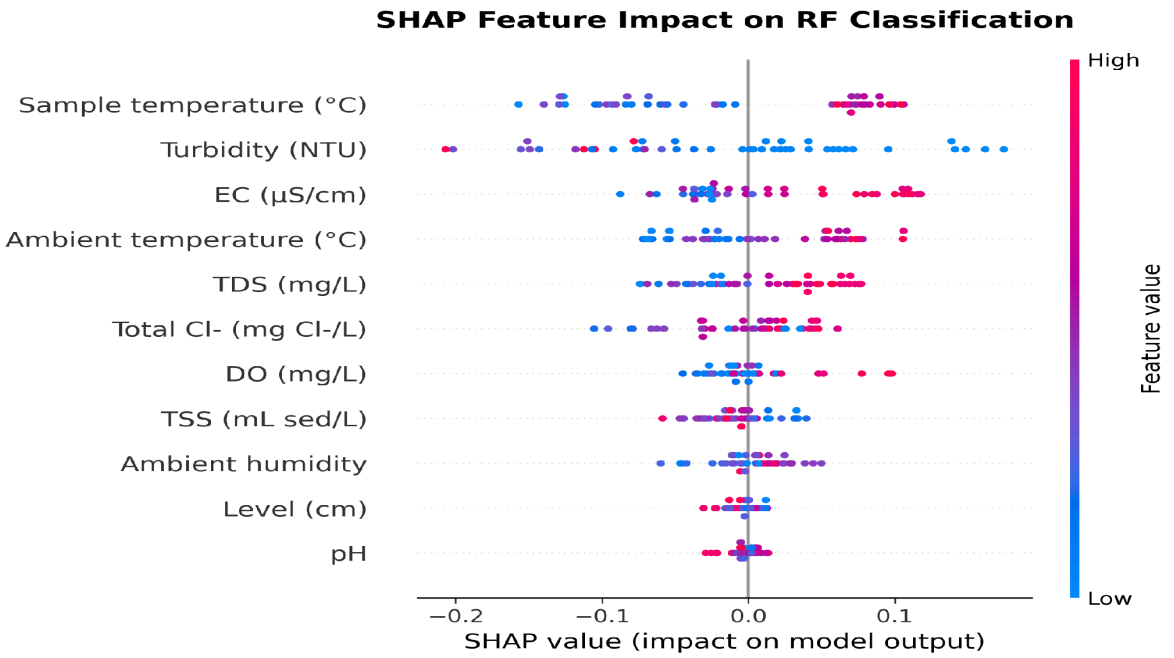


**Figure 7.** SHAP summary plot (*semidura* class, test set). Positive values push toward *semidura*; negative toward *blanda*. Red = high feature value; blue = low. High sample temperature and EC promote *semidura*; high turbidity pushes toward *blanda*; pH SHAP value is near-zero.

### 3.8 Ablation Results

To assess whether the most important features were necessary to maintain classification accuracy, an ablation study was conducted by retraining the RF classifier after removing turbidity, the top-ranked feature under both Gini and permutation importance. Despite its high importance scores, removing turbidity produced no measurable decrease in test accuracy. The combination of high attributed importance and unchanged ablation accuracy indicates that turbidity's predictive information is redundant with other features in this dataset and can be recovered through them. To examine the broader pattern of feature redundancy, the inter-feature correlation structure was analyzed.

### 3.9 Correlation Matrix Results

The Pearson correlation matrix (Figure 8) quantifies pairwise linear relationships among all eleven physicochemical predictors.

EC and TDS share a near-perfect correlation ($r = 1.00$), reflecting that they carry essentially the same physicochemical information in this dataset. This suggests that TDS may have been derived from EC using a fixed conversion factor, as is standard in multiparameter field probes, making TDS effectively a rescaled version of EC. Their redundancy may cause permutation importance to understate their independent predictive contributions (Section 3.6.2): when either is shuffled, the other preserves the signal. EC and TDS are also strongly correlated with Total $Cl^-$ ($r = 0.88$), the quantitative basis for the ionic interference described in Section 2.4.2. Turbidity and TSS are moderately correlated ($r = 0.63$), while DO and pH are correlated at $r = 0.62$. Sample temperature and ambient temperature share $r = 0.56$, pointing to partial redundancy among conceptually related sensors (interpreted in Section 4.3). These multicollinearity relationships help explain why Gini and permutation rankings diverge, and guide the reduced feature set analysis in Section 3.10.

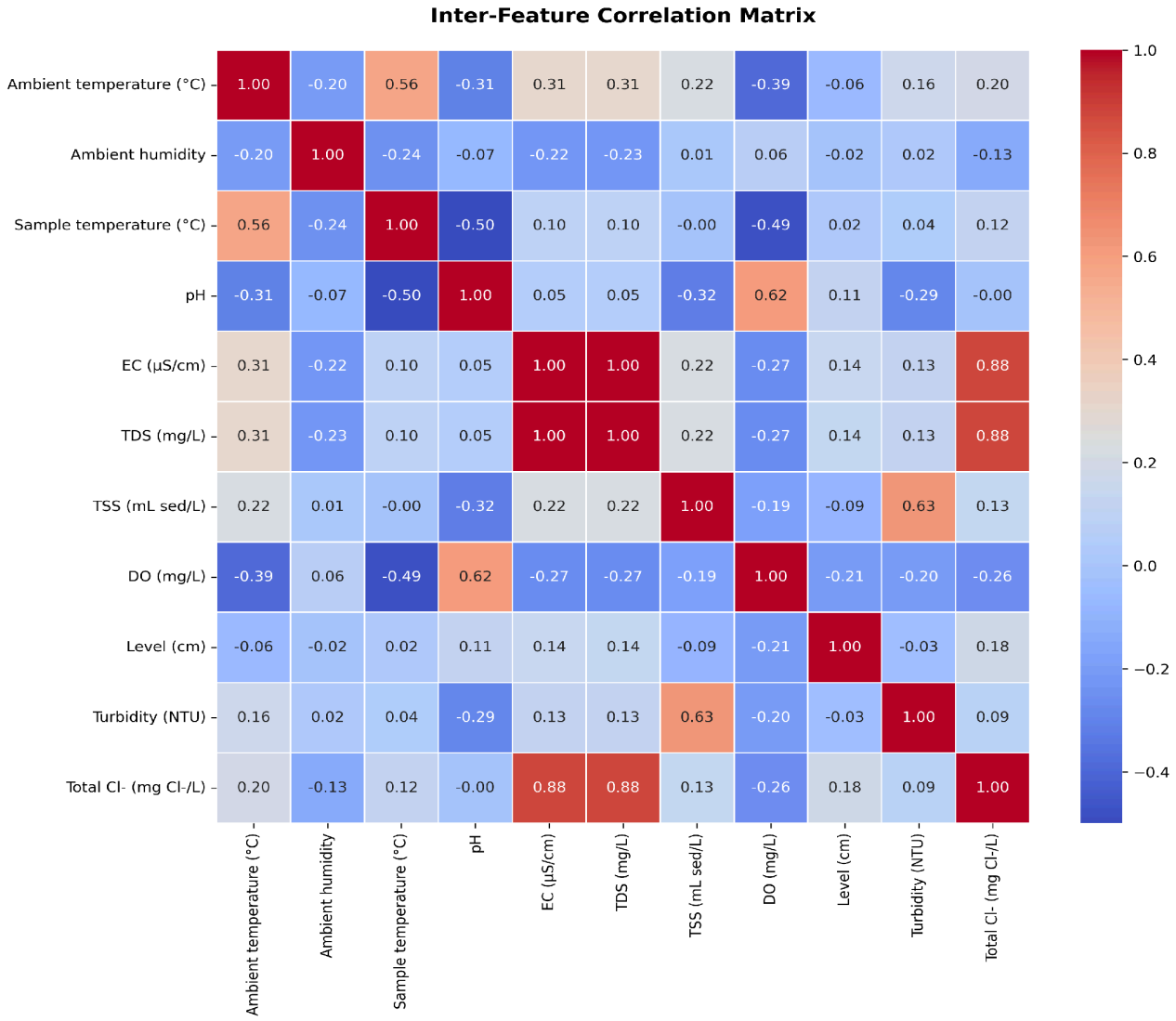


**Figure 8.** Pearson correlation matrix for all 11 features (full dataset, n = 217). EC and TDS are near-perfectly correlated (r = 1.00) and correlate strongly with Total $Cl^-$ (r = 0.88), indicating redundancy among these ionic variables. Turbidity and TSS are moderately correlated (r = 0.63), suggesting overlapping but non-identical information about suspended particulate conditions.

### 3.10 Reduced Feature Set Results

The preceding interpretability analyses describe how each feature contributes to the full model, but do not directly establish which features are indispensable in a reduced set. An exploratory feature-reduction analysis was guided by the ablation and correlation analyses.

EC was retained as the primary proxy for dissolved ionic content, consistent with its top-three ranking across all interpretability methods and its central role as the signal the model is designed to augment.

Sample temperature was retained on the basis of its high Gini rank and its SHAP-confirmed role in encoding the seasonal concentration effect. Its lower permutation rank in the full model may partly reflect overlapping information with ambient temperature. Since ambient temperature is absent from the reduced set, sample temperature may retain greater marginal utility in the reduced configuration.

Turbidity was removed because its elimination in the ablation study produced no measurable change in test accuracy, indicating that its predictive signal is recoverable through other features. The correlation matrix identified TSS as the most plausible substitute: it correlates moderately with turbidity (r = 0.63), picking up overlapping information about suspended particulate load, while showing weak correlation with EC (r = 0.22) and near-zero correlation

with sample temperature ($r \approx 0.00$). This suggests that TSS captures partially overlapping but distinct particulate information compared to turbidity, while contributing a complementary signal alongside EC and sample temperature.

TDS was not retained because its near-perfect correlation with EC ($r = 1.00$) indicates that it is effectively a conductivity-derived measurement rather than an independent feature, as discussed in Section 3.9. Total $Cl^-$ was likewise excluded because its strong correlation with EC ($r = 0.88$) indicates substantial overlap in the dissolved-ion signal. Although DO ranked second under permutation importance, this was not corroborated by Gini importance or SHAP rankings (seventh in both). Combined with its moderate correlation with pH ($r = 0.62$) and the absence of a direct hydrochemical link to hardness as provided by EC, TSS, and sample temperature, DO was considered less suitable as a feature candidate.

This three-channel configuration, an ionic proxy (EC), a particulate and dilution proxy (TSS), and a thermal proxy (sample temperature), was evaluated on the held-out test set. It achieved 95.45% accuracy, matching the full 11-feature model, while reducing the number of modeled input features by approximately 73%. Figure 9 presents the ROC curves for the full 11-feature model (AUC = 0.9888) and the reduced 3-feature model (AUC = 0.9516). While both models achieve identical 95.45% test accuracy, the minimal model shows a modest AUC reduction of 0.0372, indicating somewhat lower overall discrimination across classification thresholds.

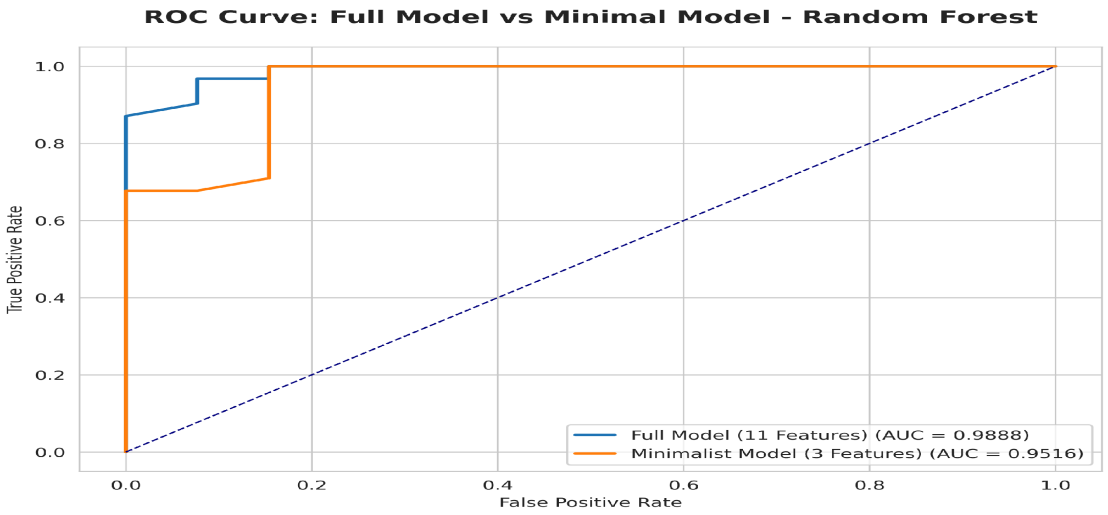


**Figure 9.** ROC curves: full 11-feature model (AUC = 0.9888) vs. 3-feature minimal model (AUC = 0.9516) using EC, TSS, and sample temperature. Both achieve 95.45% accuracy. The AUC gap of 0.0372 points to lower overall discrimination across classification thresholds for the latter.

## 4. Discussion

### 4.1 Performance beyond the EC Baseline

The RF model achieved 95.45% accuracy, a 20.45 percentage point improvement over the EC threshold baseline of 75.00%. This supports the hypothesis that multivariate physicochemical measurements capture predictive information beyond a single conductivity-based proxy.

### 4.2 Model Behavior and Data Sufficiency

The RF classifier outperforms SVM, LR, and KNN for structural reasons that go beyond model complexity. Water hardness in this catchment is governed by the interaction of three complementary geochemical signals: dissolved ionic load (captured by EC and TDS),

hydrological dilution state (captured by turbidity and TSS), and thermal-seasonal context (captured by sample temperature). The class boundary appears to depend on threshold-like changes in several variables rather than on one smooth linear trend. RF's grid-like decision cuts (Breiman, 2001) are well-matched to this structure. Each recursive tree split separates observations according to thresholds in the measured variables, allowing RF to approximate complex, threshold-like interactions among ionic, hydrological, and thermal factors. LR is limited to a single linear boundary and cannot capture these step-like transitions. KNN is sensitive to noise and less effective with increasing dimensionality as distances become less discriminative. SVM's relatively lower performance likely reflects the sensitivity of the RBF kernel to untuned hyperparameters, combined with the effect of the 71:29 class imbalance on the learned decision boundary.

Data sufficiency is indicated by the learning curve approaching an asymptotic performance plateau. Although the curve has not fully flattened, the rate of improvement decreases substantially as the maximum training size is approached, suggesting diminishing gains before full saturation. These results are consistent with the model learning stable predictive structure from this dataset, though cross-catchment generalization cannot be confirmed without external validation. The persistent gap between training and CV accuracy points to some degree of overfitting. The mean CV estimate of 89.84% is therefore a more conservative estimate of expected performance than the 95.45% accuracy observed on the held-out test set.

### 4.3 Physicochemical Interpretation of the Base Feature Set

Hardness is governed primarily by dissolved $Ca^{2+}$ and $Mg^{2+}$, while EC reflects contributions from all dissolved ions, conflating hardness-causing and non-hardening ionic species. The strong EC–$Cl^-$ correlation ($r = 0.88$) traces back to the substantial contribution of chloride salinity to total conductivity. Consequently, elevated EC may arise from increased chloride salinity rather than greater $Ca^{2+}$ or $Mg^{2+}$ concentrations, reinforcing why EC is an indirect rather than specific proxy for hardness.

EC is also temperature-dependent following Walden's rule. The high Gini rank of sample temperature alongside EC is a sign that the model accounts for temperature-dependent conductivity, with sample temperature potentially separating genuine ionic concentration from the viscosity-driven thermal artifact. A second, independent pathway is also plausible: sample temperature may capture a seasonal concentration effect, in which warmer months correspond to lower river flow and correspondingly higher ion concentration, pushing predictions toward *semidura* (Section 3.7). Sample temperature's correlation with ambient temperature ($r = 0.56$) shows partial redundancy bounded by the thermal inertia of the water body, which buffers sample temperature against short-term swings in ambient conditions.

Turbidity and TSS dominate the rankings because of their shared role as indicators of the particulate phase of mineral transport within the catchment. Although moderately correlated ($r = 0.63$), they measure different properties: turbidity is an optical measure, whereas TSS quantifies suspended solids gravimetrically. Consequently, changes in particle size distribution can cause the two variables to diverge, allowing each to capture overlapping but distinct aspects of the same particulate load. During high-flow events, storm runoff mobilizes suspended sediment while simultaneously diluting dissolved ions, including $Ca^{2+}$ and $Mg^{2+}$. Elevated turbidity and TSS are associated with the dilution-dominated *blanda* regime, whereas lower values correspond to the more mineral-concentrated *semidura* regime. Turbidity and TSS thus act as proxies for dilution, indicating hydrological state to the model.

This bridges the dissolved-phase ionic signal captured by EC with the suspended-phase hydrological signal captured by TSS, while turbidity provides a partially redundant measure of the same hydrological state.

pH ranks low mainly because of the buffering capacity of the carbonate-bicarbonate system described in Section 1.1, which keeps pH within a narrow range (SD = 0.29) regardless of $Ca^{2+}/Mg^{2+}$ variation. DO and pH co-vary (r = 0.62), likely because both respond to photosynthetic activity. Photosynthesis consumes dissolved $CO_2$, shifting the carbonate equilibrium and increasing pH, while simultaneously producing $O_2$, thereby increasing DO. As a result, both variables are shadowing the same underlying biological process.

### 4.4 Physicochemical Interpretation of the Minimal Feature Suite

The three features in the reduced feature set each represent a distinct and complementary information channel. EC is the primary proxy for dissolved ionic content. It tracks the combined concentration and mobility of all dissolved ions and serves as the model's direct window into total mineral load. It conflates hardness-causing $Ca^{2+}/Mg^{2+}$ with non-hardening $Na^{+}/Cl^{-}$, a limitation that TSS and sample temperature together help correct.

TSS captures the same hydrological dilution signal as turbidity (r = 0.63) without requiring an optical turbidity sensor. Its relatively weak correlation with EC (r = 0.22) and near-zero correlation with sample temperature ($r \approx 0.00$) further suggests that TSS contributes additional sediment and hydrological context for interpreting EC. It helps the model distinguish genuine $Ca^{2+}/Mg^{2+}$ enrichment from high-flow dilution events, which can produce similar EC values despite fundamentally different hydrochemical conditions.

Sample temperature provides a third source of information by capturing the thermal and seasonal conditions of the water. It helps the model account for the thermal effects on EC, since warmer water increases ionic mobility even without an actual increase in dissolved minerals. It also carries a seasonal signature: warmer, lower-flow conditions are often associated with higher hardness.

### 4.5 Practical Deployment and Sustainability Implications

The model's interpretable use of physicochemical variables supports its potential as a transparent field screening tool for sustainable water-quality assessment. Replacing routine EDTA titration with ML-based screening could eliminate reagent use for most samples. Confirmatory titration would be reserved for low-confidence cases, once an appropriate operational threshold has been established through calibration and validation, reducing overall reagent consumption. This directly enacts two of the twelve principles of green chemistry: prevention of waste at source, and elimination of laboratory chemical handling (Anastas & Warner, 1998).

In a continuous monitoring system, the model could receive repeated sensor measurements from fixed or portable probes and provide near-real-time hardness classification as water conditions change. For water utilities managing river intake systems where hardness fluctuates seasonally, this would provide more timely operational intelligence than periodic laboratory titration.

Accurate hardness classification can also provide an initial indication of scale risk in industrial water management. The Langelier Saturation Index ($LSI = Actual\ pH - Saturation\ pH$), for example, uses pH and hardness-related parameters to assess scaling tendency. However, the model provides only a preliminary binary hardness signal and should inform,

rather than directly trigger, treatment decisions such as scale inhibitor dosing; full LSI calculations require additional water-chemistry parameters.

The directional asymmetry of the model's errors is also relevant to deployment. Both misclassifications on the held-out test set were *blanda* samples predicted as *semidura*, with zero *semidura* samples assigned to *blanda*. The model thus shows a conservative tendency toward the majority class under the 71:29 imbalance, tending to overestimate rather than underestimate hardness. In water treatment applications, this is a favorable error direction: a false positive may trigger unnecessary scale inhibitor dosing (a minor cost), whereas a false negative may risk scale buildup in boilers or cooling circuits (a more serious operational failure).

The three-sensor deployment with the minimal model requires only a portable EC probe, a TSS meter, and a thermometer, enabling a potentially low-cost field monitoring setup with equipment costs below US$200. This supports cost-effective and hardware-resilient monitoring strategies in field environments where sensor fouling or failure is common. For applications requiring only a binary classification decision, the minimal model may be adequate. For applications where stronger ranking performance across thresholds is desirable, the full model's higher AUC (0.9888 vs. 0.9516) may be preferable.

## 5. Conclusion

This study demonstrates that river water hardness can be classified from routine in situ measurements with high accuracy, reducing reliance on laboratory EDTA titration. The RF classifier achieved 95.45% test accuracy and AUC = 0.9888 (five-fold CV: 89.84% ± 4.80%), outperforming both baselines as well as other ML models. Its tree-based structure effectively captures nonlinear, threshold-like relationships and feature interactions, which may help explain its advantage over LR's single linear boundary, KNN's distance-based voting, and the untuned SVM. Within the sampled catchment, the learning curve approached a plateau at the maximum training size, suggesting the dataset is close to sufficient for reliable classification; however, broader validation is still needed to confirm generalizability across catchments.

The five model analysis approaches (Gini importance, permutation importance, SHAP, ablation, and inter-feature correlation) help explain how the RF model uses the measured physicochemical variables, addressing the interpretability gap common to black-box ML approaches. The results indicate that the model captures EC–$Cl^-$ interference patterns, hydrological dilution, and thermal effects on conductivity without explicit feature engineering.

A key practical finding is that a reduced model using EC, TSS, and sample temperature achieved the same test-set accuracy as the full model while reducing the number of required sensors by 73%. These three features capture complementary geochemical signals: ionic strength, hydrological dilution state, and thermal context respectively. This enables low-cost, reagent-free hardness monitoring, subject to validation on additional catchments. The framework is relevant to industrial water management where hardness informs scale-risk assessment, and supports green chemistry principles through reduced reliance on reagent-intensive analytical methods. Together, these results show that the same analyses used to explain the model's predictions can also inform practical model design by identifying a reduced feature set that preserves the full model's test accuracy.

## 6. Limitations

Several limitations apply. The dataset derives from a single river system in Buenos Aires province over seven months; transferability to catchments with different lithology, land use, or turbidity drivers is unvalidated. The predictive role of TSS is likely catchment-specific: in systems where turbidity is driven by biological factors, such as algal blooms, rather than mineral suspension, its relationship with hydrological dilution may not hold. Compounding this, the random stratified split does not account for the longitudinal campaign structure of the dataset. Campaign-aware or site-aware CV, such as leaving one campaign or site out, would provide a better assessment of generalization across sampling conditions.

With only 217 samples, the dataset is modest and the high test AUC should therefore be interpreted cautiously. Additionally, the same 44-sample test set was used both to guide feature selection via the ablation study, and to evaluate the resulting minimal model, creating a risk of selection bias. Cross-validated feature selection and ablation would provide more stable estimates of feature-set performance. Other combinations of minimal feature sets warrant a thorough and systematic exploration.

Hyperparameter optimization (e.g., RF depth and number of trees, SVM $C$ and $\gamma$, KNN $k$) was not performed. Randomized- or grid-search could potentially improve performance, so the model comparison is limited to the selected configurations rather than the maximum achievable performance of each model.

Only binary classification was attempted; a model predicting continuous hardness values would provide richer information for LSI calculation.

Although in-situ TSS sensors are available, site-specific calibration may limit their practicality relative to turbidity sensors, making sensor choice dependent on local availability and the characteristics of suspended particles at the monitoring site.

## 7. Future Work

Future research should evaluate model transferability across catchments with substantially different hydrochemical and turbidity regimes, including highly complex systems such as the Ganga and Yamuna rivers in India. The Yamuna, in particular, receives substantial industrial discharge from tanneries, textile mills, and chemical plants, carrying high loads of $Na^+$ and $Cl^-$ that would substantially decouple EC from hardness, while turbidity is driven by a mixture of industrial particulates and agricultural sediment with different optical-to-chemical characteristics. Testing the RF in such an environment would provide a rigorous assessment of whether the learned feature interactions generalize beyond the present catchment.

Several further directions are priorities. First, integrating the lightweight three-feature model with continuous, low-power sensor platforms could enable real-time hardness screening in the field. Bluetooth-enabled probes could support GPS-tagged records and cloud upload, enabling reagent-free field monitoring under changing hydrological conditions and across multiple sampling locations. Second, continuous hardness regression would enable direct LSI calculation from probe measurements, providing richer quantitative output for industrial water management, when combined with the additional parameters LSI requires. Third, a delta-temperature feature (sample temperature minus ambient temperature) should be tested as an engineered predictor potentially capturing thermal differences associated with recent weather conditions or sampling context. Lastly, gradient-boosting methods such as XGBoost and LightGBM should be evaluated as additional benchmarks.

## 8. Acknowledgments

The author gratefully acknowledges Natanael Ferran for making the source dataset publicly available on Kaggle (Ferran, 2024). The author also thanks his father, Rajiv Ratnam, whose encouragement inspired both this study and the perseverance to bring it to completion.

## 9. References

Anastas, P. T., & Warner, J. C. (1998). Green chemistry: Theory and practice. Oxford University Press. https://global.oup.com/academic/product/green-chemistry-9780198506980

Baird, R. B., Eaton, A. D., & Rice, E. W. (Eds.). (2017). Standard methods for the examination of water and wastewater (23rd ed.). American Public Health Association, American Water Works Association, & Water Environment Federation. https://www.standardmethods.org/doi/book/10.2105/smww.2882

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Ferran, N. (2024). River water parameters [Dataset]. Kaggle. https://www.kaggle.com/datasets/natanaelferran/river-water-parameters

Haghiabi, A. H., Nasrolahi, A. H., & Parsaie, A. (2018). Water quality prediction using machine learning methods. Water Quality Research Journal, 53(1), 3–13. https://doi.org/10.2166/wqrj.2018.025

Hem, J. D. (1985). Study and interpretation of the chemical characteristics of natural water (3rd ed.). U.S. Geological Survey Water Supply Paper 2254. https://pubs.usgs.gov/wsp/wsp2254/pdf/wsp2254a.pdf

Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data, 6(4), 1–21. https://doi.org/10.1145/2382577.2382579

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1705.07874

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://dl.acm.org/doi/10.5555/1953048.2078195

Ratnam, A. R. (2026). waterhardnessML [Computer software]. GitHub. https://github.com/adityaratnam09/waterhardnessML

World Health Organization. (2011). Hardness in drinking-water: Background document for development of WHO guidelines for drinking-water quality (WHO/HSE/WSH/10.01/10/Rev/1). https://cdn.who.int/media/docs/default-source/wash-documents/wash-chemicals/hardness-bd.pdf

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