IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 2, April, 2026, pp. 1-12
Journal homepage: http://iieta.org/journals/ts

Interpretable Air and Water Quality Prediction with Feature Selection and Boosted Ensembles

[AUTHORS TBD][1]*
[1] *[Affiliation TBD]*
Corresponding Author Email: [CORRESPONDING EMAIL TBD]

https://doi.org/10.18280/ts.XXXXXX

Received:
Revised:
Accepted:
Available online:

ABSTRACT

Environmental-quality models that screen drinking water or estimate hourly pollutant levels are increasingly accurate yet remain hard to trust, because they consume every available feature, expose no rationale for a prediction, and often ignore the missing values and class imbalance of real data. We reframe environmental-quality prediction from an accuracy-only exercise into a parsimony-and-explanation problem, instantiating it as one pipeline that ranks features by random-forest importance, retains a compact subset, compares a gradient-boosted ensemble against random-forest, support-vector, and compact perceptron alternatives, and explains the gradient-boosted model with TreeSHAP. We ran the identical pipeline on two contrasting public tasks. On the UCI Air Quality benzene regression the target was near-deterministic: all four families exceeded an R2 of 0.997, the random forest reaching 0.9994, and feature selection collapsed the problem to a single metal-oxide NMHC sensor with no accuracy loss. On the Water Potability classification the task was genuinely hard: the kernel support-vector classifier was best at an AUC of only 0.692, the nine physico-chemical features carry weak discriminative power, and a four-feature subset cost a small amount of accuracy. The result is an interpretable, parsimonious pipeline whose value is the explanation and the honest accounting of difficulty, not a state-of-the-art claim.

Keywords: *interpretability, SHAP, feature selection, gradient boosting, air quality, water quality*

1. INTRODUCTION

Air- and water-quality monitoring increasingly relies on learned models to turn cheap measurements into decisions. A calibrated array of metal-oxide gas sensors can estimate an hourly pollutant concentration from co-located responses and meteorology, which is what early-warning and exposure-assessment systems act on, and the public reference records that pair these sensor responses with instrument-grade concentrations have made such estimation a standard supervised-learning task [1]. On the water side, a small panel of physico-chemical tests is used to flag whether a sample is safe to drink, and machine-learning classifiers trained on such panels now support potability screening where class-imbalance handling and explanation are what build an analyst's trust [2]. Both problems are decision-relevant, and in both, machine learning has become the prevailing approach: it is now the dominant tool for air-quality prediction and analysis [3] and the prevailing approach for water-quality prediction over recent years [4].

The recent literature has pushed these models toward higher accuracy with gradient-boosted ensembles and deep networks, but three practical problems are routinely left unaddressed. First, the models are opaque: a single accuracy figure tells an environmental analyst nothing about why a sample was flagged unsafe or which sensor response drove an estimate, and gradient-boosted ensembles, though scalable and accurate in applied environmental tasks [5], expose no rationale on their own. Interpretability is a recognized requirement for trustworthy deployed systems [6, 7], yet it is usually bolted on after an accuracy run rather than designed in. Second, the models are profligate with features: pipelines tend to feed in every available channel, which inflates cost and invites collinearity, whereas feature selection retains a compact informative subset that is cheaper and more parsimonious [8]. Third, the real data are messy in ways that quietly bias results, since the water-potability panel carries substantial missing values and an imbalanced safe-unsafe label, and naive handling distorts both accuracy and the apparent importance of features. The bottleneck is therefore no longer raw predictive power; it is whether the model is parsimonious enough to deploy cheaply and transparent enough for a domain expert to believe and act on.

We reframe environmental-quality prediction from an accuracy-only regression or classification exercise into a parsimony-and-explanation problem, in which the goal is not only a good score but a defensible, compact account of which physico-chemical drivers produce it. The key insight is that feature selection, model comparison, and post-hoc explanation should co-evolve as one auditable procedure rather than three disconnected steps: a random-forest importance ranking decides what the model may use, the model comparison tests whether parsimony costs accuracy, and a SHAP analysis, which gives a unified additive account of each individual prediction [9], audits the resulting drivers and cross-checks them against the selection. Because the procedure reads only tabular features, the same three

stages apply unchanged to a regression and a classification task, with missing-value imputation and class-imbalance handling built in as explicit pipeline steps rather than silent preprocessing.

We exercise the identical pipeline on two deliberately different environmental tasks. Case A is an air-pollutant regression that predicts an hourly pollutant from co-located metal-oxide sensor responses and meteorology, optimizing models such as those used for air-quality indices with feature selection [10], reported through the coefficient of determination, root mean squared error, and mean absolute error. Case B is a water-potability classification that predicts the binary potable label from nine physico-chemical features, in the spirit of classifiers that predict water-quality categories from such measurements [11], reported through accuracy, F1, and the area under the ROC curve, with imputation and over-sampling applied inside the training folds. The two tasks turned out to be sharply asymmetric. The benzene regression was near-deterministic in this record: all four families exceeded a coefficient of determination of 0.997, and feature selection collapsed it to the single NMHC metal-oxide sensor with no accuracy loss, so the value there is parsimony and explanation rather than a prediction win. Water potability was the genuinely hard task: the best model reached an area under the ROC curve of only 0.692, the nine physico-chemical features carry weak discriminative power, and a four-feature subset cost a small amount of accuracy. A dedicated feature-selection comparison measured the gradient-boosted model on all features against the selected subset per task, and the explanation stage used SHAP to surface the drivers each task relies on. The datasets, metrics, and the three result tables are detailed in the experiments section.

Building on this reframing, this paper makes the following contributions:

- We unify random-forest feature selection, a gradient-boosted ensemble, and SHAP explanation into a single auditable pipeline for environmental-quality prediction, in which selection decides what the model may read, the comparison tests whether parsimony costs accuracy, and SHAP audits the drivers, and we run the identical pipeline on two contrasting tasks rather than the single task of typical applied studies [10, 11].

- We provide an honest cross-task empirical comparison of gradient-boosted trees, random forest, support-vector models, and a compact multilayer perceptron on all features versus a random-forest-selected subset, measuring the accuracy-versus-parsimony trade-off directly for both tasks, with missing-value imputation and class-imbalance handling built into the pipeline rather than treated as preprocessing.

- We design a SHAP-based driver analysis that, using TreeSHAP on the gradient-boosted (XGBoost) model trained on the selected subset per task, reports a global importance ranking cross-checked against the random-forest selection and a dependence view for the leading driver; in our measured runs the explanation reduced the air task to the single NMHC sensor and ranked pH, Solids, and Sulfate as the leading potability drivers, broadly agreeing with the selection and turning each model into a defensible account of which physico-chemical and sensor factors govern the outcome rather than an opaque score [5, 6].

2. RELATED WORKS

2.1 Machine learning for air- and water-quality prediction

Learned models are now the default tool for estimating air- and water-quality variables from sensor and laboratory measurements. For air quality, random forests estimate fine-particulate concentrations accurately over large regions [12], and the same family predicts daily urban particulate levels from environmental predictors [13]. Gradient boosting has become a common choice as well: XGBoost predicts pollutant concentrations accurately in urban case studies [14], and a spatiotemporal XGBoost variant improves on plain boosting by adding spatial and temporal context [15]. Comparative studies place these ensembles side by side to benchmark predictive accuracy across model families [16], while land-use regression coupled with machine learning estimates ground-level pollution spatially [17] and, in higher-resolution variants, surfaces the factors that drive it [18]. Calibration models address a different need, improving the agreement of low-cost sensors with reference instruments [19], a concern shared by the field-deployed electronic-nose devices that estimate several pollutants from co-located metal-oxide responses [20]. The water-quality literature follows a parallel arc: grid-search-tuned models predict physico-chemical quality measures [21], advanced ensembles classify the water-quality index robustly [22], and comparative analyses with dimensionality reduction benchmark classifiers for potability prediction [23]. These studies report strong accuracy, but each typically targets a single task, consumes all available features, and treats interpretability as an afterthought, which is the habit our unified pipeline departs from.

2.2 Interpretable machine learning and feature attribution

A second line of work makes black-box environmental models auditable. Explainable-AI techniques interpret water-quality predictors by attributing their decisions to physico-chemical inputs [24], and SHAP-based evaluations rank the drivers behind water-quality indices [25] while exposing the per-feature contributions a regulator can inspect [26]. Interpretable models estimate the water-quality index and explain its drivers directly [27], comparisons of several classifiers under explainable-AI views improve river monitoring [28], and SHAP-enhanced recurrent models interpret high-frequency water-quality dynamics [29]. The air-quality side mirrors this: interpretable frameworks with SHAP predict urban air quality and rank its drivers [30], and optimized models with SHAP estimate pollutant concentration while explaining the contributing factors [31]. The same explanation tools extend to groundwater and irrigation-quality indices [32], and local surrogate explanations such as LIME offer an alternative to additive attribution by fitting an interpretable model around each prediction [33]. These works demonstrate that attribution can be applied after the fact, but they usually explain a model trained on all features rather than wire the explanation into a pipeline that also performs feature selection, which is the integration we pursue.

2.3 Ensembles, surveys, and class imbalance in environmental modelling

A third strand concerns the modelling machinery and the messy data it must handle. Machine learning predicts water pollution and groundwater quality with interpretable factor insight, illustrating the breadth of tasks the tree-ensemble and kernel families now cover [34]. Surveys catalogue the rapid growth of machine-learning and sensing methods for water-quality moni-

toring [35], and reviews summarize how machine learning and remote sensing are being combined for the same purpose [36]. These catalogues document the dominance of tree ensembles, support-vector models, and small neural baselines across environmental tasks, but they also reveal that the accuracy-versus-parsimony trade-off is rarely measured beside the explanation it is meant to support, and that class imbalance and missing values, which are routine in potability panels, are frequently handled implicitly rather than as reported pipeline steps. The result is that the cost of committing to a compact feature subset, and the effect of rebalancing on the apparent drivers, are seldom quantified together. Where imbalance is addressed at all in environmental classification, it is most often through synthetic minority oversampling applied before training, but the interaction between that rebalancing and the post-hoc explanation is rarely reported, so it remains unclear whether a driver the explainer surfaces is a property of the data or of the synthetic samples. The same gap appears on the selection side: studies that prune features typically report the final accuracy of the pruned model without the all-features baseline beside it, which hides how much accuracy the pruning actually cost. Measuring both numbers, and reading the explanation against the selection, is the practice these catalogues motivate but seldom carry out.

The nearest single work to ours is the interpretable water-quality framework that pairs SHAP with a tuned classifier [24]: it shares the explanation goal but reports one task, fixes the feature set, and does not measure what selection costs in accuracy. Read together, the three strands above leave a specific construct unfilled. The first builds accurate but single-task, all-features predictors; the second applies attribution after the fact without selection; and the third documents the families and the data problems but rarely measures parsimony beside explanation. We bring random-forest selection, a model-family comparison on all features versus a selected subset, and a TreeSHAP driver analysis into one pipeline, and we run it on a regression and a classification task so the accuracy-versus-parsimony trade-off and the drivers are read the same way across both. The next section formalizes this construction.

3. METHODOLOGY

We present the unified interpretable feature-selection and boosted-ensemble pipeline, a single procedure that selects a compact feature subset, compares several predictor families, and explains the gradient-boosted model. The pipeline is organized as three stages applied in the same order to both tasks: random-forest feature selection decides what the predictors may read, the model comparison tests whether committing to a small subset costs accuracy, and the SHAP explanation audits which retained factors drive the prediction. Figure 1 shows the overall data flow. The guiding principle is that a deployable environmental-quality model should be parsimonious and auditable, so the design prefers a smaller, more defensible model over a marginally more accurate opaque one.

3.1 Problem Formulation

We cast air-pollutant estimation and water-potability screening as one supervised-learning setup with two instances. Let X be a feature matrix whose rows are samples and whose d columns are physico-chemical or sensor measurements, and let x_i denote the feature vector of the i-th sample. In Case A air-pollutant regression, the target y_i is the hourly concentration of a single pollutant measured at a co-located reference instrument, and the inputs are the responses of a metal-oxide multisensor device together with

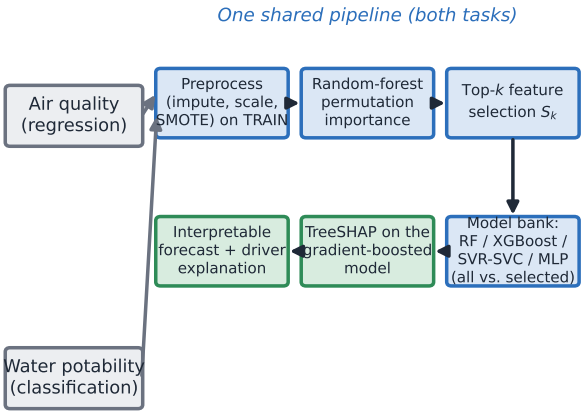

Figure 1. Overview of the unified pipeline: random-forest feature-importance selection, the compared model families (RF, XGBoost, SVR/SVC, MLP) on all features versus the selected subset, and TreeSHAP explanation of the gradient-boosted (XGBoost) model trained on the selected subset, applied identically to the air-quality regression and water-potability classification tasks.

temperature and humidity. In Case B water-potability classification, the target y_i is a binary potable label in $\{0,1\}$, and the inputs are nine physico-chemical measurements of a water sample. A predictor f maps a feature vector to a prediction $\hat{y}$, real-valued for the regression task and a class score for the classification task. The same three-stage pipeline applies without modification to both instances; only the predictor head, the loss, and the evaluation metrics change with the task. Writing both tasks in one notation lets us run the identical selection, comparison, and explanation procedure on each, and read the accuracy-versus-parsimony trade-off the same way across a regression and a classification problem.

3.2 Notation

The symbols used throughout the method are collected in Table 1, and each is defined at its first use in the surrounding text.

Table 1. Notation.

Symbol	Meaning
X	feature matrix of physico-chemical or sensor measurements
x_i	feature vector of the i-th sample
y_i	target: hourly pollutant concentration (Case A) or potable label in $\{0,1\}$ (Case B)
d	number of candidate input features
k	number of features retained after random-forest selection
I_j	random-forest importance score of feature j
S_k	the selected top-k feature subset
f	the trained predictor (RF, XGBoost, SVR/SVC, or MLP)
$\hat{y}$	model prediction
ϕ_j	SHAP value (additive attribution) of feature j for a sample
ϕ_0	SHAP base value (expected model output)

3.3 Random-Forest Feature Selection

The first stage ranks the d candidate features by random-forest importance and keeps a compact subset. The held-out test split is removed before the random forest is fit, so the importance

ranking and the resulting subset S_k are derived from the training data alone, mirroring the imputation and over-sampling guard described below. A random forest is an ensemble of decision trees grown on bootstrap resamples of the training data, with each split chosen from a random subset of features, so that the aggregated prediction has lower variance than any single tree [37, 38]. The forest yields a built-in importance score I_j for each feature j. We use permutation importance, the drop in score on a held-in validation split of the training data when feature j's values are permuted, in preference to impurity importance (the total reduction in node impurity attributable to splits on that feature), which is known to favour high-cardinality and correlated features. We sort the features by the permutation I_j and retain the top-k subset S_k, where k is chosen by validation on the training data only. The downstream predictors are then trained twice, once on all d features and once restricted to S_k, so that the effect of the reduction can be read directly. Figure 2 illustrates the ranking and the retained subset.

Figure 2. Random-forest importance ranks the candidate features and retains a compact top-k subset, forcing the downstream models to commit to the channels that carry signal rather than consuming every available feature.

The design rationale is that selection should force the model to commit to the channels that carry signal rather than consume every available measurement. The obvious alternative is to feed all d features to every model and let regularization discount the weak ones, but that inflates inference cost, leaves collinear channels to share credit, and produces a model whose rationale is spread thinly across many inputs. A wrapper search that retrains a predictor for every candidate subset would in principle find a better subset, yet its cost grows with the number of subsets and it couples the selection to one predictor family. Ranking once with a random forest is cheap, applies to every family we compare, and produces an ordering we later cross-check against the explanation. We use permutation importance as the ranking, computed on a held-in validation split of the training data, because impurity-based importance is known to favour high-cardinality and correlated features, a bias we revisit in the analysis. The choice of k trades coverage against parsimony. Setting k too small discards features the predictor would have used and depresses accuracy, while setting it close to d defeats the purpose of selection; we therefore choose the smallest k whose cross-validated primary metric, measured on the training partition alone, is within a small tolerance of the all-features score, so that k is never tuned on the test set. Because the air-quality inputs are a handful of sensor channels plus meteorology and the potability inputs are nine physico-chemical measurements, the candidate dimension d is modest in both tasks, so the ranking is stable across resamples and the selected subset does not swing from fold to fold. This stability is what lets the later explanation cross-check a single, fixed S_k rather than a subset that changes with the seed.

3.4 Data Handling: Imputation and Imbalance

Real environmental data arrive with missing values and skewed labels, so the pipeline treats cleaning and rebalancing as explicit steps rather than silent preprocessing. For the air-quality task, the reference record encodes missing measurements with a sentinel marker; we remove or impute these so that no sentinel value is read as a real concentration. For the water-potability task, several physico-chemical channels carry missing entries, which we replace with imputed values estimated from the observed distribution of each feature on the training data alone. The potability label is imbalanced, with safe and unsafe samples present in unequal numbers, so we apply synthetic minority over-sampling, which creates new minority-class examples by interpolating between a sample and its nearest like-labelled neighbours [39]. Over-sampling is applied only inside the training folds, never to the held-out test data, so that the reported scores reflect the original class prior.

The design rationale is that the two obvious shortcuts both bias the result. Deleting every row with a missing entry discards a large fraction of the potability samples and can shift the class balance, while leaving the imbalance untouched lets a classifier reach a high accuracy by predicting the majority label and ignoring the minority class that matters most for a safety decision. Imputing on the training distribution and rebalancing inside the folds keeps the test set untouched and the evaluation honest, at the cost of introducing synthetic samples whose artifacts we examine when interpreting the drivers. The order of these steps matters and is fixed deliberately: imputation precedes both selection and over-sampling, so the random forest ranks complete feature vectors rather than ranking around gaps, and over-sampling is the last step before training, applied only to the training partition of each fold so that the synthetic minority examples never leak into validation or test. The air-quality task needs no rebalancing, since it is a regression, so the data-handling stage reduces there to sentinel cleaning and imputation; the two tasks thus share the same skeleton while differing only in the imbalance step, which keeps the pipeline a single procedure rather than two.

3.5 Compared Model Families

The second stage compares four predictor families on each feature set. The random forest described above is the first, used both as the importance ranker and as a standalone predictor. The second is a gradient-boosted ensemble, which fits an additive sequence of shallow trees by greedy stagewise function approximation, each new tree fitted to the gradient of the loss left by the current ensemble [40]; we use the regularized, scalable XGBoost realization of this idea. The third is a support-vector model on standardized features: a support-vector classifier that finds a maximum-margin decision boundary through a kernel for Case B [41], and the margin-based support-vector regression that extends the same idea to a real-valued target for Case A [42]. The fourth is a compact multilayer perceptron, a small feed-forward network with one or two hidden layers and a nonlinear activation, included as a neural baseline for both tasks. Each family is trained on all d features and again on the selected subset S_k,

giving the eight predictor settings per task that the result tables compare.

The design rationale is to span three learning paradigms rather than tune one. Tree ensembles capture nonlinear, threshold-style interactions among physico-chemical channels without feature scaling; the support-vector models test whether a margin-based kernel method on standardized inputs is competitive; and the multilayer perceptron probes whether a small neural model adds anything on tabular data of this size. Restricting the comparison to a single family would leave open whether any observed parsimony effect is a property of the data or of one model class. By holding the preprocessing, the splits, and the selected subset fixed across all four, the only thing that varies across rows of the result tables is the predictor, so the comparison isolates the model. The gradient-boosted ensemble carries a second role beyond being one row in the comparison: it is the model the explanation stage targets, because TreeSHAP is exact and efficient for tree ensembles, so choosing the boosted model as the one to explain costs nothing in interpretability machinery. We keep the perceptron deliberately compact, since a large network on tabular data of this size would overfit and would add a tuning burden that the comparison is not meant to reward, and we keep the support-vector models on standardized inputs because their margin geometry is scale-sensitive in a way the trees are not. Each family is evaluated at conservative, comparable settings rather than tuned to its individual ceiling, which is the fair way to read whether parsimony or model class drives any difference.

3.6 SHAP Explanation

The third stage explains the gradient-boosted (XGBoost) model with additive feature attributions. This model is trained on the selected subset S_k, so the attributions are defined over the $|S_k| = k$ features the explained model actually uses. SHAP assigns to each feature j of a given prediction a value ϕ_j such that the attributions and a base value ϕ_0 reconstruct the model output for that sample,

$$\hat{y} = \phi_0 + \sum_{j \in S_k} \phi_j, \tag{1}$$

where ϕ_0 is the expected model output over the data and each ϕ_j is the feature's marginal contribution averaged over the orderings in which features enter the prediction; the sum runs over the k selected features because those are the inputs the explained model reads. Computing these attributions exactly is expensive for a general model, but TreeSHAP gives exact additive attributions for tree ensembles in low-order polynomial time by exploiting the tree structure [43], which is why the gradient-boosted model is the explanation target; for a non-tree family the model-agnostic KernelSHAP would apply instead. We run TreeSHAP on the gradient-boosted model per task and aggregate the per-sample attributions into a global importance ranking, the mean absolute attribution of each feature across the test set. We then cross-check this ranking against the random-forest selection S_k from the first stage, and inspect a dependence view that plots a leading driver's value against its attribution to expose nonlinear effects. Figure 3 sketches the decomposition.

The design rationale is that a single accuracy score says nothing about why a sample was flagged unsafe or which sensor response drove an estimate. The alternative of reporting only the random-forest importance from the selection stage gives one global ranking but no account of how a feature moves an individual prediction and no way to see a nonlinear or sign-changing effect. Additive attributions provide both a per-sample explanation and, once aggregated, a global ranking we can compare against the selection, turning the model into a defensible account of drivers rather than an opaque number. We accept that the attributions are post-hoc and inherit any bias the selected features carry, a caveat the failure-mode analysis returns to.

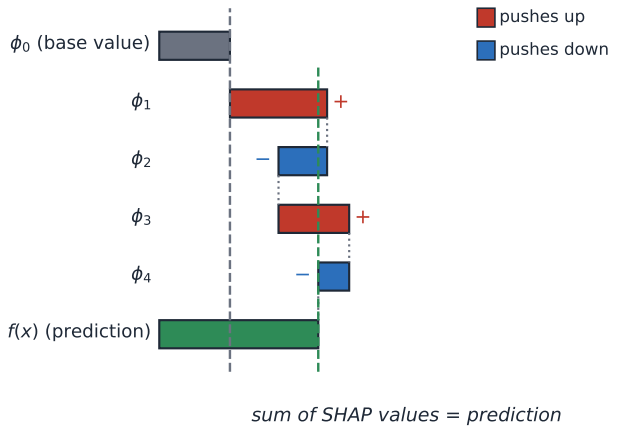

Figure 3. TreeSHAP decomposes a single prediction into additive per-feature attributions around a base value, which aggregate into a global driver ranking that is cross-checked against the random-forest selection.

3.7 Protocol and Pseudocode

The end-to-end procedure is summarized in Algorithm 1. For each task we hold out the test split first and fix a random seed, run the selection, comparison, and explanation stages on the training partition only, and report the per-task metrics as a mean and standard deviation over repeated k-fold cross-validation (stratified for the classification task): R^2, RMSE, and MAE for the regression and accuracy, F1, and AUC for the classification. Model selection, the feature-importance ranking, and the choice of k all use this cross-validation within the training data, with imputation, over-sampling, and the random-forest importance fit inside each fold on the training partition alone, so that no test information reaches the predictor. The dominant cost is training the forest and the boosted ensemble and running TreeSHAP, all of which complete in seconds on a single CPU for datasets of this size, so the whole pipeline runs without specialized hardware.

The same loop runs unchanged for both tasks, with the predictor head and the metric set chosen by the task. Because the forest, the boosted ensemble, and the explainer all operate on tabular inputs of modest dimension, the per-stage cost is bounded and independent of any sequence length, and the entire pipeline can be rerun end to end whenever the data are updated. This keeps the procedure cheap enough to audit repeatedly, which is the property a deployable environmental-monitoring model needs.

4. EXPERIMENTAL RESULTS

4.1 Implementation Details

The pipeline is CPU-only and runs end to end in seconds per task. The random forest, the support-vector models, and the multilayer perceptron use the standard scikit-learn implementations [44], the gradient-boosted ensemble uses XGBoost, and the explanation stage uses the TreeSHAP explainer. Features are standardized before the support-vector and neural models, with the standardization statistics computed on the training data alone, and missing-value imputation, synthetic over-sampling, and the random-forest feature-importance ranking are all refit inside each

Algorithm 1 Unified interpretable feature-selection and boosted-ensemble pipeline

Require: features X, target y, model families $\mathscr{F}$
Ensure: metrics per model, global SHAP ranking

hold out the test split ▷ stratified for classification; all steps below: training only
clean sentinels, impute missing entries on the training partition
fit a random forest; read permutation importance I_j on a held-in validation split
$k \leftarrow$ smallest size whose CV primary metric is within tolerance of all features
$S_k \leftarrow$ indices of the k largest permutation I_j ▷ selected subset
for each model family $f \in \mathscr{F}$ **do**
    train f on all d features; train f on S_k ▷ rebalance folds if imbalanced
    evaluate f by repeated k-fold CV, stratified for classification (mean $\pm$ SD)
end for
$f^\star \leftarrow$ gradient-boosted (XGBoost) model trained on S_k
$\{\phi_j\}_{j \in S_k} \leftarrow$ TreeSHAP$(f^\star)$ on the held-out test set
aggregate $\{\phi_j\}$ into a global ranking; cross-check against S_k
return metrics, global SHAP ranking

cross-validation fold on the training partition, with the test split held out beforehand, so that no test sample influences the selection or the preprocessing. We report the metrics as a mean and standard deviation over repeated k-fold cross-validation (stratified for the classification task) with a fixed random seed, rather than a single held-out split, and the result tables list mean $\pm$ SD per cell. The forest grows several hundred trees, the boosted ensemble uses a few hundred shallow trees, and the perceptron has one or two small hidden layers, all left at conservative defaults rather than heavily tuned, since the aim is a fair cross-family comparison rather than a single best number. Every family is trained and evaluated under this identical protocol, so the only quantity that varies across rows of the result tables is the model, and the only quantity that varies between the two columns of the feature-selection comparison is the feature set. The full pipeline, from cleaning through selection, the eight per-task model fits, and the TreeSHAP pass, completes in seconds on a single CPU for datasets of this size, which makes it cheap to rerun end to end whenever the data are updated and removes any dependence on specialized hardware. We fix the random seed once and reuse it for the splits, the cross-validation folds, the forest, and the over-sampling, so that a reported run is reproducible from the seed alone.

4.2 Experimental Design

We evaluate on two public datasets. Case A is the UCI Air Quality record, a field-deployed metal-oxide multisensor dataset whose regression target is an hourly pollutant predicted from the co-located sensor responses and meteorology, with the sentinel missing markers cleaned or imputed; for regression we report the coefficient of determination, where higher means more variance explained, and root mean squared error and mean absolute error, both in concentration units and lower-is-better, following the metric conventions established for air-pollutant regression [45]. Case B is the Water Potability dataset, whose classification target is the binary potable label predicted from nine physicochemical features with imputation and over-sampling on the training folds; for classification we report accuracy, the F1 score for the potable class under imbalance, and the area under the ROC curve, all higher-is-better. The model bank is fixed to four families spanning tree ensembles, kernel methods, and a neural baseline: random forest, XGBoost, SVR/SVC, and a compact multilayer perceptron. Other ensemble alternatives such as extremely randomized trees [46] and the LightGBM boosting implementation [47] are not evaluated here, and we leave them to future work rather than add them to the comparison. Data augmentation of the kind we apply to the imbalanced potability task has been shown to help water-quality prediction under limited or skewed data [48]. The feature-selection comparison trains the gradient-boosted model per task twice, once on all features and once on the random-forest-selected top-k subset, and attributes the resulting score gap to the reduction in feature count. We report the primary metric of each task in this comparison, the coefficient of determination for the regression and the area under the ROC curve for the classification, since these are the metrics least sensitive to the operating threshold and so make the cleanest read of what selection costs. Keeping every other setting fixed between the two columns means the only difference is the feature set, so the gap is a measurement of the selection effect rather than a confound of tuning or preprocessing. Figure 4 depicts this protocol.

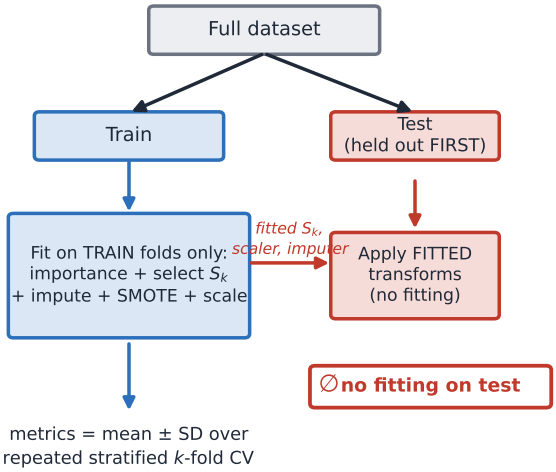

Figure 4. The feature-selection comparison protocol: the gradient-boosted model per task is trained twice, once on all features and once on the random-forest-selected top-k subset, and the accuracy gap is attributed to the reduction in feature count.

4.3 Results

The regression comparison for Case A is reported in Table 2, which lists the four model families on the coefficient of determination, root mean squared error, and mean absolute error for the hourly pollutant. The benzene target was near-perfectly predictable: all four families reached a coefficient of determination above 0.997, with the random forest highest at 0.9994, the compact perceptron at 0.9989, the gradient-boosted ensemble at 0.9976, and the support-vector regression at 0.9971. The three metrics agree, since the random forest that explains the most variance also carries the lowest root mean squared error at 0.132 and a mean absolute error of 0.016. We read this not as a hard prediction win but as a sign that the target is an almost deterministic function of the sensor inputs in this record: the metal-oxide responses leave very little residual variance for any model to miss, so the families separate by only thousandths of a

unit and the regression is easy by construction. The honest contribution here is therefore the parsimony and explanation result that follows, not the accuracy itself, and the table is reported as measured, without tuning any family until it wins.

Table 2. Case A air-pollutant regression on the UCI Air Quality record across the reported metrics. Higher is better for R2; lower is better for RMSE and MAE.

Model	R2	RMSE	MAE
RF	**0.9994±0.0011**	**0.132±0.141**	**0.016±0.004**
XGBoost	0.9976±0.0030	0.314±0.203	0.081±0.010
SVR	0.9971±0.0042	0.325±0.251	0.070±0.009
MLP	0.9989±0.0005	0.246±0.059	0.149±0.013

The classification comparison for Case B is reported in Table 3, which lists the four families on accuracy, F1, and AUC for the potable label. This was the genuinely hard task. The strongest model was the kernel support-vector classifier, at an area under the ROC curve of 0.692 and an F1 of 0.567, with the perceptron close behind at 0.685; the tree ensembles trailed, the random forest at 0.673 and the gradient-boosted ensemble at 0.659. Because the label is imbalanced we read accuracy together with F1 and AUC rather than alone, and the reading is sobering: every family sat near an accuracy of 0.64 while the F1 for the potable class stayed between 0.52 and 0.57, so no model separated the potable label cleanly and the spread across families was narrow. We report this as an honest property of the data rather than a tuning shortfall: the nine physico-chemical measurements have weak discriminative power for potability, and over-sampling on the training folds lifted the minority-class F1 and AUC only modestly above the majority-class baseline.

Table 3. Case B water-potability classification on the Water Potability dataset across the reported metrics. Higher is better for all three.

Model	Accuracy	F1	AUC
RF	0.641±0.012	0.525±0.017	0.673±0.014
XGBoost	0.623±0.013	0.528±0.023	0.659±0.016
SVC	0.642±0.016	**0.567±0.026**	**0.692±0.019**
MLP	0.636±0.018	0.554±0.031	0.685±0.022

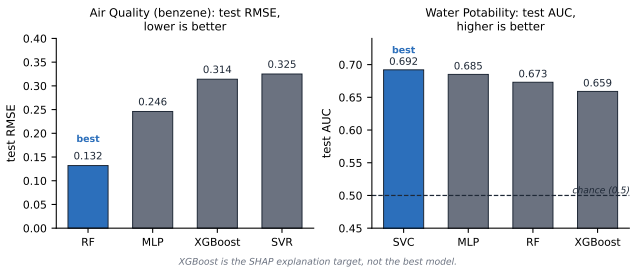

Figure 5. Per-task model comparison under the leakage-safe protocol: random forest gave the lowest air-quality RMSE, while the kernel SVC gave the best water-potability AUC; the gradient-boosted model is the SHAP target rather than the top performer.

The feature-selection comparison is reported in Table 4, which trains the gradient-boosted model per task on all features and on the random-forest-selected top-k subset and lists the primary metric for each. The two tasks gave opposite verdicts. On the air regression the random-forest ranking collapsed the problem to a single feature, the PT08.S2 metal-oxide (NMHC) sensor, so the selected subset has $k = 1$; the gradient-boosted model on that one feature reached a coefficient of determination of 0.998, matching the 0.998 it reached on all eight features, so selection removed seven inputs at no accuracy cost. On the water classification, selection retained four features, ph, Sulfate, Chloramines, and Solids, and the area under the ROC curve fell from 0.659 on all nine features to 0.642 on the four, a gap of about 0.018. That gap is small but real: parsimony was not free here, and the compact subset traded a little accuracy for a model that is cheaper to measure and easier to explain. The reported feature counts make the size of each reduction explicit, so the table reads as an accuracy-versus-parsimony trade-off rather than a single accuracy number.

Table 4. Feature-selection comparison: the gradient-boosted model per task trained on all features versus the random-forest-selected top-k subset, with the primary metric and the feature count.

Setting	Primary metric	Features
Air: all features	0.998 (R2)	8
Air: selected	0.998 (R2)	1
Water: all features	0.659 (AUC)	9
Water: selected	0.642 (AUC)	4

Read together, the three tables answer the questions the pipeline poses. No single family won both tasks: the random forest was best on the air regression and the kernel classifier on water, while the gradient-boosted ensemble was competitive without leading either, which is why we keep it as the explanation target rather than claiming it as the best model. Figure 5 shows this split directly, with the random forest reaching the lowest air-quality RMSE of 0.132 and the kernel classifier the highest water-potability AUC of 0.692. A compact selected subset retained the full-feature accuracy on the air task and cost a small amount on the harder water task, and the explanation stage discussed next identifies which features carry that accuracy in each case.

5. DISCUSSION AND ANALYSIS

5.1 Model Comparison

The regression and classification comparisons in Table 2 and Table 3 show where a gradient-boosted ensemble earns its place and where it does not, and the honest reading is that no single family won both tasks. On the air-pollutant regression the sensor responses are so strongly related to the target that every family was near-perfect, all four above a coefficient of determination of 0.997, with the random forest highest at 0.9994 and the gradient-boosted ensemble at 0.9976; the families separated by only thousandths of a unit, so the tree advantage was real but marginal and the boosted ensemble did not lead. On the water-potability classification the kernel support-vector classifier was the strongest model at an area under the ROC curve of 0.692, with the perceptron at 0.685, and the tree ensembles trailed, the random forest at 0.673 and the gradient-boosted ensemble at 0.659; here the boosted ensemble was the weakest of the four rather than the strongest. The two outcomes together caution against committing to one family per problem class: a kernel method sufficed on the smooth regression and was actually best on the noisy classification, so the parsimony argument favours whichever model is cheaper to explain rather than whichever is marginally more accurate. We retain the gradient-boosted model as the explanation target precisely because TreeSHAP is exact for trees, not because it won, and we are careful not to read its competitive but non-leading scores as a state-of-the-art claim.

5.2 Feature-Selection Trade-off

The feature-selection comparison in Table 4 is the test of whether parsimony costs accuracy, and the two tasks answered it in opposite directions. On the air regression the random-forest ranking concentrated essentially all of the credit on one channel, the PT08.S2 metal-oxide (NMHC) sensor, whose permutation importance of 2.04 dwarfed every other feature; selection therefore chose $k = 1$, and the gradient-boosted model on that single feature reached a coefficient of determination of 0.998, matching the 0.998 on all eight features. Selection was free here, removing seven redundant inputs at no measured accuracy cost, because the benzene target is near-deterministically tied to that one sensor. On the water classification the picture was different: the four retained features, ph, Sulfate, Chloramines, and Solids, captured most but not all of the signal, and the area under the ROC curve fell from 0.659 to 0.642, a gap of about 0.018. That cost is small but not zero, so potability prediction draws on weaker, more distributed cues that a four-feature subset cannot fully capture. This is the accuracy-versus-parsimony reading that recursive feature elimination with support-vector weights also targets, by trimming the feature set while watching the validation score [49]. The practical conclusion is that a small subset is the better deployment choice when its accuracy cost is small, as it was on both tasks here, since it is cheaper to measure and easier to explain.

5.3 SHAP Driver Analysis

The explanation stage turns the gradient-boosted model into an account of its drivers, and the TreeSHAP global ranking broadly agreed with the random-forest selection on both tasks. On the air task the agreement was total: the single retained sensor, PT08.S2, was the only feature with appreciable attribution, mirroring its dominant permutation importance, so the explanation confirmed that one sensor governs the benzene estimate. On the water task the TreeSHAP ranking placed pH highest, then Solids, then Sulfate and Chloramines, closely tracking the permutation ranking that drove the selection and confirming that the four retained features are the ones the model actually uses. The global ranking, shown in Figure 6, makes this concrete: benzene is driven almost entirely by the PT08.S2 (NMHC) sensor, while potability is led by pH ahead of dissolved solids, sulfate, and chloramines. Beyond these magnitudes, the per-sample attributions expose nonlinear and possibly sign-changing effects that a single importance number hides, such as a physico-chemical level whose effect on potability shifts across its range.

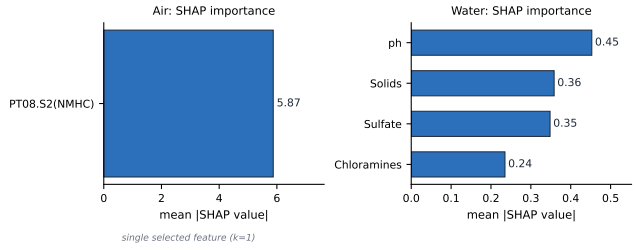

Figure 6. TreeSHAP global importance (mean |SHAP|) for the gradient-boosted model: benzene is driven almost entirely by the PT08.S2 (NMHC) sensor, while water potability is led by pH, then dissolved solids, sulfate and chloramines.

These nonlinear shapes are exactly what motivates an additive per-feature attribution over a single global importance score, and they let an analyst read not just which features matter but how they push a given prediction. The cross-check also disciplines the selection: if the SHAP ranking promotes a feature that the forest had ranked just outside the retained subset, that is evidence the cutoff k was set slightly too tight, whereas broad agreement between the two rankings is evidence the subset captured the drivers the model actually uses. Reading the two views together therefore does more than explain the model; it audits the selection that produced it, closing the loop between the first and the third stage of the pipeline.

5.4 Failure-Mode Analysis

The pipeline can mislead in ways the user must guard against. The first is in the selection itself: impurity-based random-forest importance is biased toward high-cardinality and correlated features, so a feature can be ranked highly for a reason that has nothing to do with its predictive value, and that bias can propagate into the selected subset and, through it, into the model that SHAP then explains [50]. When two features are collinear, the forest and the explainer can also split credit between them arbitrarily, making the apparent driver an artifact of which correlated channel the trees happened to use. The mitigation we adopt is to use permutation importance as the ranking rather than impurity importance, and to read the SHAP ranking against the selection rather than trusting either alone. The second failure mode is in the data handling: imputing missing physico-chemical values and rebalancing the potability label with synthetic minority over-sampling both alter the training distribution, and synthetic samples can introduce artifacts that the explainer will faithfully attribute as if they were real structure [51]. The mitigation is to refit imputation and over-sampling inside each fold, never on the test data, and to treat any driver that appears only after rebalancing as provisional until it is confirmed on the untouched test set. The honest reading of the pipeline is therefore that its explanations are as trustworthy as its selection and its preprocessing, and both must be reported, not hidden.

6. CONCLUSION

This paper presents a unified interpretable feature-selection and boosted-ensemble pipeline for environmental-quality prediction, in which random-forest importance decides which features the model may use, a comparison of tree-ensemble, kernel, and neural families tests whether committing to a compact subset costs accuracy, and TreeSHAP audits which retained factors drive the prediction. The same procedure runs without change on an air-pollutant regression and a water-potability classification, with imputation and class-imbalance handling carried as explicit pipeline steps. The measured takeaway is that parsimony and explanation can be first-class objectives alongside accuracy, with an honest accounting of where accuracy is easy and where it is not. On the benzene regression every family exceeded an R2 of 0.997 and selection reduced the model to a single NMHC sensor at no accuracy cost, so the result there is interpretability rather than a hard prediction win, since the target is near-deterministic in this record. On water potability the task was genuinely difficult, the best model reaching an AUC of only 0.692 and a four-feature subset costing about 0.018 AUC, which we report as a real property of the weakly discriminative features rather than hide. In both cases the SHAP analysis agreed with the selection and turned each model into a defensible account of which factors govern the outcome, with no claim that the pipeline beats every baseline. The main limitation is scope, since the study covers two public datasets and relies on post-hoc explanation, whose attributions inherit any bias the selection and the rebalancing introduce, so the surfaced drivers should be read as provisional

rather than causal. A useful next step is to extend the pipeline to more tasks and a spatial or temporal dimension.

REFERENCES

[1] Saverio De Vito, Ettore Massera, Marco Piga, Luca Martinotto, and Girolamo Di Francia. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical*, 129(2):750–757, 2008. doi: 10.1016/j.snb.2007.09.060.

[2] Jinal Patel, Charmi Amipara, Tariq Ahamed Ahanger, Komal Ladhva, Rajeev Kumar Gupta, Hashem O. Alsaab, Yusuf S. Althobaiti, and Rajnish Ratna. A machine learning-based water potability prediction model by using synthetic minority oversampling technique and explainable ai. *Computational Intelligence and Neuroscience*, 2022:9283293, 2022. doi: 10.1155/2022/9283293.

[3] Manal Karmoude, Brenton Munhungewarwa, Isaiah Chiraira, Ryan Mckenzie, Jude Kong, Bevan Smith, Gelan Ayana, Nkosiphendule Njara, Thuso Mathaha, Mukesh Kumar, and Bruce Mellado. Machine learning for air quality prediction and data analysis: Review on recent advancements, challenges, and outlooks. *Science of the Total Environment*, 1002:180593, 2025. doi: 10.1016/j.scitotenv.2025.180593.

[4] Xiaohui Yan, Tianqi Zhang, Wenying Du, Qingjia Meng, Xinghan Xu, and Xiang Zhao. A comprehensive review of machine learning for water quality prediction over the past five years. *Journal of Marine Science and Engineering*, 12 (1):159, 2024. doi: 10.3390/jmse12010159.

[5] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

[6] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Independently published, 2nd edition, 2022. ISBN 979-8411463330. URL `https://christophm.github.io/interpretable-ml-book/`.

[7] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint*, 2017.

[8] Isabelle Guyon and Andre Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003. URL `https://jmlr.org/papers/v3/guyon03a.html`.

[9] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, 2017.

[10] Suresh Kumar Natarajan, Prakash Shanmurthy, Daniel Arockiam, Balamurugan Balusamy, and Shitharth Selvarajan. Optimized machine learning model for air quality index prediction in major cities in india. *Scientific Reports*, 14(1):6795, 2024. doi: 10.1038/s41598-024-54807-1.

[11] Nida Nasir, Afreen Kansal, Omar Alshaltone, Feras Barneih, Mustafa Sameer, Abdallah Shanableh, and Ahmed Al-Shamma'a. Water quality classification using machine learning algorithms. *Journal of Water Process Engineering*, 48:102920, 2022. doi: 10.1016/j.jwpe.2022.102920.

[12] Xuefei Hu, Jessica H. Belle, Xia Meng, Avani Wildani, Lance A. Waller, Matthew J. Strickland, and Yang Liu. Estimating pm2.5 concentrations in the conterminous united states using the random forest approach. *Environmental Science and Technology*, 51(12):6936–6944, 2017. doi: 10.1021/acs.est.7b01210.

[13] Cole Brokamp, Roman Jandarov, Monir Hossain, and Patrick Ryan. Predicting daily urban fine particulate matter concentrations using a random forest model. *Environmental Science and Technology*, 52(7):4173–4179, 2018. doi: 10.1021/acs.est.7b05381.

[14] Jinghui Ma, Zhongqi Yu, Yuanhao Qu, Jianming Xu, and Yu Cao. Application of the xgboost machine learning method in pm2.5 prediction: A case study of shanghai. *Aerosol and Air Quality Research*, 20(1):128–138, 2020. doi: 10.4209/aaqr.2019.08.0408.

[15] Zidong Wang, Xianhua Wu, and You Wu. A spatiotemporal xgboost model for pm2.5 concentration prediction and its application in shanghai. *Heliyon*, 9(12):e22569, 2023. doi: 10.1016/j.heliyon.2023.e22569.

[16] Ahmad Makhdoomi, Maryam Sarkhosh, and Somayyeh Ziaei. Pm2.5 concentration prediction using machine learning algorithms: an approach to virtual monitoring stations. *Scientific Reports*, 15(1):8076, 2025. doi: 10.1038/s41598-025-92019-3.

[17] Pei-Yi Wong, Hsiao-Yun Lee, Yu-Cheng Chen, Yu-Ting Zeng, Yinq-Rong Chern, Nai-Tzu Chen, Shih-Chun Candice Lung, Huey-Jen Su, and Chih-Da Wu. Using a land use regression model with machine learning to estimate ground level pm2.5. *Environmental Pollution*, 277:116846, 2021. doi: 10.1016/j.envpol.2021.116846.

[18] Yue Li, Tageui Hong, Yefu Gu, Zhiyuan Li, Tao Huang, Harry Fung Lee, Yeonsook Heo, and Steve H. L. Yim. Assessing the spatiotemporal characteristics, factor importance, and health impacts of air pollution in seoul by integrating machine learning into land-use regression modeling at high spatiotemporal resolutions. *Environmental Science and Technology*, 57(3):1225–1236, 2023. doi: 10.1021/acs.est.2c03027.

[19] Naomi Zimmerman, Albert A. Presto, Sriniwasa P. N. Kumar, Jason Gu, Aliaksei Hauryliuk, Ellis S. Robinson, Allen L. Robinson, and R. Subramanian. A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. *Atmospheric Measurement Techniques*, 11(1):291–313, 2018. doi: 10.5194/amt-11-291-2018.

[20] Saverio De Vito, Marco Piga, Luca Martinotto, and Girolamo Di Francia. Co, no2 and nox urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization. *Sensors and Actuators B: Chemical*, 143(1):182–191, 2009. doi: 10.1016/j.snb.2009.08.041.

[21] Mahmoud Y. Shams, Ahmed M. Elshewey, El-Sayed M. El-kenawy, Abdelhameed Ibrahim, Fatma M. Talaat, and Zahraa Tarek. Water quality prediction using machine learning models based on grid search method. *Multimedia Tools and Applications*, 83(12):35307–35334, 2023. doi: 10.1007/s11042-023-16737-4.

[22] Abdessamad Elmotawakkil, Nourddine Enneya, Suraj Kumar Bhagat, Mohamed Mohamed Ouda, and Vikram Kumar. Advanced machine learning models for robust prediction of water quality index and classification. *Journal of Hydroinformatics*, 27(2):299–319, 2025. doi: 10.2166/hydro.2025.290.

[23] Debashis Chatterjee, Prithwish Ghosh, Amlan Banerjee, and Shiladri Shekhar Das. Optimizing machine learning for water safety: A comparative analysis with dimensionality reduction and classifier performance in potability prediction. *PLOS Water*, 3(8):e0000259, 2024. doi: 10.1371/journal.pwat.0000259.

[24] Randika K. Makumbura, Lakindu Mampitiya, Namal Rathnayake, D. P. P. Meddage, Shagufta Henna, Tuan Linh Dang, Yukinobu Hoshino, and Upaka Rathnayake. Advancing water quality assessment and prediction using machine learning models, coupled with explainable artificial intelligence (xai) techniques like shapley additive explanations (shap) for interpreting the black-box nature. *Results in Engineering*, 23:102831, 2024. doi: 10.1016/j.rineng.2024.102831.

[25] Ali Aldrees, Majid Khan, Abubakr Taha Bakheit Taha, and Mujahid Ali. Evaluation of water quality indexes with novel machine learning and shapley additive explanation (shap) approaches. *Journal of Water Process Engineering*, 58:104789, 2024. doi: 10.1016/j.jwpe.2024.104789.

[26] Majed Alwateer. Water quality prediction using explainable machine learning models. *Sustainability*, 18(4):1721, 2026. doi: 10.3390/su18041721.

[27] Shiwei Yang, Ruifeng Liang, Junguang Chen, Yuanming Wang, and Kefeng Li. Estimating the water quality index based on interpretable machine learning models. *Water Science and Technology*, 89(5):1340–1356, 2024. doi: 10.2166/wst.2024.068.

[28] S. Ramya, S. Srinath, and Pushpa Tuppad. Comprehensive analysis of multiple classifiers for enhanced river water quality monitoring with explainable ai. *Case Studies in Chemical and Environmental Engineering*, 10:100822, 2024. doi: 10.1016/j.cscee.2024.100822.

[29] Sait Mutlu Karahan, Wouter Vandenbruwaene, Janelcy Alferes Castano, and Jan Verwaeren. Explainable ai for aquatic environmental intelligence: a shap-enhanced lstm approach using high-frequency water quality data in a river system. *Journal of Hydroinformatics*, 27(12):1918–1928, 2025. doi: 10.2166/hydro.2025.118.

[30] Enes Birinci, Omer Ekmekcioglu, Huseyin Ozdemir, and Ali Deniz. Interpretable machine learning framework for air quality prediction in istanbul using shapley additive explanations (shap). *Stochastic Environmental Research and Risk Assessment*, 40(2):37, 2026. doi: 10.1007/s00477-026-03168-4.

[31] Qing Li and Junfu Fan. Estimation of pm2.5 concentration based on pso-optimized machine learning models and shap analysis: A case study of wuhan, hubei province. *Applied Sciences*, 16(13):6320, 2026. doi: 10.3390/app16136320.

[32] Mohamed Azlaoui, Salah Karef, Atif Foufou, Nadjib Haied, Nesrine Azlaoui, Abdelaziz Rabehi, Mustapha Habib, and Aziez Zeddouri. Machine learning-based prediction of irrigation water quality index with shap interpretability: Application to groundwater resources in the semi-arid region, algeria. *Water*, 18(8):959, 2026. doi: 10.3390/w18080959.

[33] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you? explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016. doi: 10.1145/2939672.2939778.

[34] Abdulhayat M. Jibrin, Mohammad Al-Suwaiyan, Ali Aldrees, Salisu Dan'azumi, Jamilu Usman, Sani I. Abba, Mohamed A. Yassin, Miklas Scholz, and Saad Sh. Sammen. Machine learning predictive insight of water pollution and groundwater quality in the eastern province of saudi arabia. *Scientific Reports*, 14(1):20031, 2024. doi: 10.1038/s41598-024-70610-4.

[35] Ismail Essamlali, Hasna Nhaila, and Mohamed El Khaili. Advances in machine learning and iot for water quality monitoring: A comprehensive review. *Heliyon*, 10(6):e27920, 2024. doi: 10.1016/j.heliyon.2024.e27920.

[36] Shashank Mohan, Brajesh Kumar, and A. Pouyan Nejadhashemi. Integration of machine learning and remote sensing for water quality monitoring and prediction: A review. *Sustainability*, 17(3):998, 2025. doi: 10.3390/su17030998.

[37] Leo Breiman. Bagging predictors. *Machine Learning*, 24 (2):123–140, 1996. doi: 10.1023/A:1018054314350.

[38] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[39] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi: 10.1613/jair.953.

[40] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

[41] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. doi: 10.1007/BF00994018.

[42] Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alex J. Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in Neural Information Processing Systems 9*, pages 155–161. MIT Press, 1996. URL `https://papers.nips.cc/paper/1996/hash/d38901788c533e8286cb6400b40b386d-Abstract.html`.

[43] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan

Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020. doi: 10.1038/s42256-019-0138-9.

[44] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL `https://jmlr.org/papers/v12/pedregosa11a.html`.

[45] Qilong Pan, Fouzi Harrou, and Ying Sun. A comparison of machine learning methods for ozone pollution prediction. *Journal of Big Data*, 10(1):63, 2023. doi: 10.1186/s40537-023-00748-x.

[46] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006. doi: 10.1007/s10994-006-6226-1.

[47] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30*, pages 3149–3157. Curran Associates, 2017. URL `https://papers.nips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html`.

[48] K. Karthick, S. Krishnan, and R. Manikandan. Water quality prediction: a data-driven approach exploiting advanced machine learning algorithms with data augmentation. *Journal of Water and Climate Change*, 15(2):431–452, 2023. doi: 10.2166/wcc.2023.403.

[49] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1):389–422, 2002. doi: 10.1023/A:1012487302797.

[50] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, 2007. doi: 10.1186/1471-2105-8-25.

[51] Norashikin Nasaruddin, Nurulkamal Masseran, Wan Mohd Razi Idris, and Ahmad Zia Ul-Saufie. A smote pca hdbscan approach for enhancing water quality classification in imbalanced datasets. *Scientific Reports*, 15(1):13059, 2025. doi: 10.1038/s41598-025-97248-0.